

**METHODS FOR IDENTIFYING RISK OF BREAST CANCER  
AND TREATMENTS THEREOF**

Related Patent Applications

[0001] This patent application claims the benefit of provisional patent application no. 60/429,136 filed November 25, 2002 and provisional patent application no. 60/490,234 filed July 24, 2003, having attorney docket number 524593004100 and 524593004101, respectively. Each of these provisional patent applications names Richard B. Roth *et al.* as inventors and is hereby incorporated herein by reference in its entirety, including all drawings and cited publications and documents. Also incorporated by reference are patent applications concurrently filed on November 25, 2003, the day this application is filed, entitled "Methods for identifying risk of breast cancer and treatments thereof," naming Richard B. Roth *et al.* as inventors and bearing attorney docket numbers 524592006600, 524592006700, 524592006800, 524592007000, 524592007100 and 524592007200. In addition, incorporated by reference is a concurrently filed patent application naming Matthew R. Nelson as an inventor, entitled "Disease risk prediction with associated single nucleotide polymorphisms," having attorney docket number 524593006400.

Field of the Invention

[0002] The invention relates to genetic methods for identifying risk of breast cancer and treatments that specifically target the disease.

Background

[0003] Breast cancer is the third most common cancer, and the most common cancer in women, as well as a cause of disability, psychological trauma, and economic loss. Breast cancer is the second most common cause of cancer death in women in the United States, in particular for women between the ages of 15 and 54, and the leading cause of cancer-related death (Forbes, *Seminars in Oncology*, vol.24(1), Suppl 1, 1997: pp.S1-20-S1-35). Indirect effects of the disease also contribute to the mortality from breast cancer including consequences of advanced disease, such as metastases to the bone or brain. Complications arising from bone marrow suppression, radiation fibrosis and neutropenic sepsis, collateral effects from therapeutic interventions, such as surgery, radiation, chemotherapy, or bone marrow transplantation-also contribute to the morbidity and mortality from this disease.

[0004] While the pathogenesis of breast cancer is unclear, transformation of normal breast epithelium to a malignant phenotype may be the result of genetic factors, especially in women under thirty (Miki, *et al.*, *Science*, 266: 66-71 (1994)). However, it is likely that other, non-genetic factors also

have a significant effect on the etiology of the disease. Regardless of its origin, breast cancer morbidity increases significantly if it is not detected early in its progression. Thus, considerable efforts have focused on the elucidation of early cellular events surrounding transformation in breast tissue. Such efforts have led to the identification of several potential breast cancer markers. For example, alleles of the *BRCA1* and *BRCA2* genes have been linked to hereditary and early-onset breast cancer (Wooster, *et al.*, Science, 265: 2088-2090 (1994)). However, *BRCA1* is limited as a cancer marker because *BRCA1* mutations fail to account for the majority of breast cancers (Ford, *et al.*, British J. Cancer, 72: 805-812 (1995)). Similarly, the *BRCA2* gene, which has been linked to forms of hereditary breast cancer, accounts for only a small portion of total breast cancer cases.

#### Summary

[0005] It has been discovered that certain polymorphic variations in human genomic DNA are associated with the occurrence of breast cancer. In particular, polymorphic variants in loci containing *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1/FLJ20625/LOC220074* (hereafter referred to as “*NUMA1*”), and *HT014/LOC148902/LYPLA2/GALE* (hereafter referred to as “*GALE*”) regions in human genomic DNA have been associated with risk of breast cancer.

[0006] Thus, featured herein are methods for identifying a subject at risk of breast cancer and/or a risk of breast cancer in a subject, which comprises detecting the presence or absence of one or more polymorphic variations associated with breast cancer in genomic regions described herein in a human nucleic acid sample. In an embodiment, two or more polymorphic variations are detected in two or more regions selected from the group consisting of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE*. In certain embodiments, 3 or more, or 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 or more polymorphic variants are detected. In specific embodiments, the group of polymorphic variants detected comprise or consist of polymorphic variants in *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE*, such as position 44247 in SEQ ID NO: 1 (*ICAM*), position 36424 in SEQ ID NO: 2 (*MAPK10*), position 48563 in SEQ ID NO: 3 (*KIAA0861*), position 49002 in SEQ ID NO: 4 (*NUMA1*) and position 174 in SEQ ID NO: 5 (*GALE*), for example.

[0007] Also featured are nucleic acids that include one or more polymorphic variations associated with the occurrence of breast cancer, as well as polypeptides encoded by these nucleic acids. Further, provided is a method for identifying a subject at risk of breast cancer and then prescribing to the subject a breast cancer detection procedure, prevention procedure and/or a treatment procedure. In addition, provided are methods for identifying candidate therapeutic molecules for treating breast cancer and related disorders, as well as methods for treating breast cancer in a subject by diagnosing breast cancer in

the subject and treating the subject with a suitable treatment, such as administering a therapeutic molecule.

[0008] Also provided are compositions comprising a breast cancer cell and/or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid with a RNAi, siRNA, antisense DNA or RNA, or ribozyme nucleic acid designed from a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence. In an embodiment, the nucleic acid is designed from a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence that includes one or more breast cancer associated polymorphic variations, and in some instances, specifically interacts with such a nucleotide sequence. Further, provided are arrays of nucleic acids bound to a solid surface, in which one or more nucleic acid molecules of the array have a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence, or a fragment or substantially identical nucleic acid thereof, or a complementary nucleic acid of the foregoing. Featured also are compositions comprising a breast cancer cell and/or a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, with an antibody that specifically binds to the polypeptide. In an embodiment, the antibody specifically binds to an epitope in the polypeptide that includes a non-synonymous amino acid modification associated with breast cancer (e.g., results in an amino acid substitution in the encoded polypeptide associated with breast cancer). In certain embodiments, the antibody specifically binds to an epitope that comprises a proline at amino acid position 352 or an alanine at amino acid position 348 in an *ICAM5* polypeptide.

#### Brief Description of the Figures

[0009] Figures 1A-1Y show a genomic nucleotide sequence for an *ICAM* region encoding *ICAM* 1, 4 and 5. The genomic nucleotide sequence is set forth in SEQ ID NO: 1. The following nucleotide representations are used throughout: “A” or “a” is adenosine, adenine, or adenylic acid; “C” or “c” is cytidine, cytosine, or cytidylic acid; “G” or “g” is guanosine, guanine, or guanylic acid; “T” or “t” is thymidine, thymine, or thymidylic acid; and “I” or “i” is inosine, hypoxanthine, or inosinic acid. Exons are indicated in italicized lower case type, introns are depicted in normal text lower case type, and polymorphic sites are depicted in bold upper case type. SNPs are designated by the following convention: “R” represents A or G, “M” represents A or C; “W” represents A or T; “Y” represents C or T; “S” represents C or G; “K” represents G or T; “V” represents A, C or G; “H” represents A, C, or T; “D” represents A, G, or T; “B” represents C, G, or T; and “N” represents A, G, C, or T.

[0010] Figures 2A-2U show a genomic nucleotide sequence of a *MAPK10* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 2.

[0011] Figures 3A-3NN show a genomic nucleotide sequence of a *KIAA0861* region. The genomic nucleotide sequence is set forth in SEQ ID NO: 3.

[0012] Figures 4A-4JJ show a genomic nucleotide sequence of a NUMA1/FLJ20625/LOC220074 region, referred to herein as the NUMA1 region. The genomic nucleotide sequence is set forth in SEQ ID NO: 4.

[0013] Figure 5 shows a portion of a genomic nucleotide sequence of a HT014/LOC148902/LYPLA2/GALE region, referred to herein as the GALE region. The genomic nucleotide sequence is set forth in SEQ ID NO: 5.

[0014] Figures 6A-6C show coding nucleotide sequences (cDNA) for ICAM1, ICAM4 and ICAM5, respectively. The nucleotide sequences are set forth in SEQ ID NOs: 6, 7 and 8, respectively.

[0015] Figure 7 shows a coding nucleotide sequence (cDNA) for MAPK10. The nucleotide sequence is set forth in SEQ ID NO: 9.

[0016] Figures 8A-8B show coding nucleotide sequences (cDNA) for KIAA0861. The nucleotide sequences are set forth in SEQ ID NO: 10 and 11, respectively.

[0017] Figures 9A-9B show a coding nucleotide sequence (cDNA) for NUMA1. The nucleotide sequence is set forth in SEQ ID NO: 12.

[0018] Figures 10A-10C show amino acid sequences for ICAM1, ICAM4 and ICAM5 polypeptides. The amino acid sequences are set forth in SEQ ID NOs: 13, 14 and 15, respectively.

[0019] Figure 11 shows an amino acid sequence for a MAPK10 polypeptide, which is set forth in SEQ ID NO: 16.

[0020] Figure 12 shows an amino acid sequence for a KIAA0861 polypeptide, which is set forth in SEQ ID NO: 17.

[0021] Figure 13 shows an amino acid sequence for a NUMA1 polypeptide, which is set forth in SEQ ID NO: 18.

[0022] Figure 14 shows proximal SNPs in the ICAM region in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in a region spanning positions 11851-24282, 36340-37868, 41213-41613, 70875-74228, 42407-45536, or 42407-51102 in SEQ ID NO: 1.

[0023] Figure 15 shows proximal SNPs in the MAPK10 region in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in a region spanning positions 23826-36424, 46176-62572, 4512-8467 or 13787-14355 in SEQ ID NO: 2.

[0024] Figure 16 shows proximal SNPs in the KIAA0861 region in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in a region spanning positions 42164-48563 in SEQ ID NO: 3.

[0025] Figure 17 shows proximal SNPs in the KIAA0861 region in genomic DNA. The position of each SNP on the chromosome is shown on the x-axis and the y-axis provides the negative logarithm of the p-value comparing the estimated allele to that of the control group. Also shown in the figure are exons and introns of the genes in the approximate chromosomal positions. The figure indicates that polymorphic variants associated with breast cancer are in linkage disequilibrium in a region spanning positions 174-32954, 38115-43785, 45386-52058, 52257-54411, 55303-73803 or 96470-98184 in SEQ ID NO: 4.

[0026] Figure 18 shows results of an odds-ratio meta analysis for the ICAM region.

[0027] Figure 19 shows results of an odds-ratio meta analysis for the MAPK10 region.

[0028] Figure 20 shows results of an odds-ratio meta analysis for the KIAA0861 region.

[0029] Figure 21 shows results of an odds-ratio meta analysis for the NUMA1 region.

[0030] Figure 22 shows effects of ICAM-directed siRNA on cancer cell proliferation.

#### Detailed Description

[0031] It has been discovered that polymorphic variations in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE* regions described herein are associated with an increased risk of breast cancer.

[0032] All *ICAM* proteins are type I transmembrane glycoproteins, contain 2-9 immunoglobulin-like C2-type domains, and bind to the leukocyte adhesion LFA-1 protein. The proteins are members of the intercellular adhesion molecule (*ICAM*) family. The gene *ICAM1* (intercellular adhesion molecule-1) is also known as human rhinovirus receptor, BB2, CD54, and cell surface glycoprotein P3.58. *ICAM1* has been mapped to chromosomal position 19p13.3-p13.2. *ICAM1* (CD54) typically is expressed on endothelial cells and cells of the immune system. *ICAM1* binds to integrins of type CD11a / CD18, or CD11b / CD18. *ICAM1* is also exploited by Rhinovirus as a receptor.

[0033] The gene *ICAM4* (intercellular adhesion molecule 4) is also known as the Landsteiner-Wiener blood group or LW. *ICAM4* has been mapped to 19p13.2-cen. The protein encoded by this gene is a member of the intercellular adhesion molecule (*ICAM*) family. A glutamine to arginine polymorphism in this protein is responsible for the Landsteiner-Wiener blood group system

(GLN=WB(A); ARG=WB(B)). This gene consists of 3 exons and alternative splicing generates 2 transcript variants.

[0034] The gene *ICAM5* (intercellular adhesion molecule 5) is also known as telencephalin. *ICAM5* has been mapped to 19p13.2. The protein encoded by the gene is expressed on the surface of telencephalic neurons and displays two types of adhesion activity, homophilic binding between neurons and heterophilic binding between neurons and leukocytes. It may be a critical component in neuron-microglial cell interactions in the course of normal development or as part of neurodegenerative diseases.

[0035] The gene *MAPK10* also is known as JNK3, JNK3A, PRKM10, p493F12, FLJ12099, p54bSAPK MAP kinase, c-Jun kinase 3, JNK3 alpha protein kinase, c-Jun N-terminal kinase 3, stress activated protein kinase JNK3, stress activated protein kinase beta. *MAPK10* has been mapped to chromosomal position 4q22.1-q23. The protein encoded by this gene is a member of the MAP kinase family. MAP kinases act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. This protein is a neuronal-specific form of c-Jun N-terminal kinases (JNKs). Through its phosphorylation and nuclear localization, this kinase plays regulatory roles in the signaling pathways during neuronal apoptosis. Beta-arrestin 2, a receptor-regulated MAP kinase scaffold protein, is found to interact with, and stimulate the phosphorylation of this kinase by MAP kinase kinase 4 (MKK4). Cyclin-dependent kinase 5 can phosphorylate, and inhibit the activity of this kinase, which may be important in preventing neuronal apoptosis. Four alternatively spliced transcript variants encoding distinct isoforms have been reported.

[0036] The gene *KIAA0861* is a Rho family guanine-nucleotide exchange factor. *KIAA0861* has been mapped to chromosomal position 3q27.3. *KIAA0861* is a Rho family nucleotide exchange factor homolog that modulates the activity of Rho family GTPases, which control numerous cell functions, including cell growth, adhesion, movement and shape. RhoC GTPase is overexpressed in invasive (inflammatory) breast cancers.

[0037] The gene *FLJ20625* has been mapped to chromosomal position 11q13.3. The gene encoding LOC220074 also is known as Hypothetical 55.1 kDa protein F09G8.5 in chromosome III and has been mapped to chromosomal position 11q13.3.

[0038] The gene *HT014* has been mapped to chromosomal position 1p36.11. The gene *LYPLA2* (lysophospholipase II) also is known as APT-2, DJ886K2.4 and acyl-protein thioesterase and has been mapped to chromosomal position 1p36.12-p35.1. Lysophospholipases are enzymes that act on biological membranes to regulate the multifunctional lysophospholipids. There are alternatively spliced transcript variants described for this gene but the full length nature is not known yet.

[0039] The gene *GALE* (galactose-4-epimerase, UDP-) also is known as galactowaldenase UDP galactose-4-epimerase and has been mapped to chromosomal position 1p36-p35. This gene encodes UDP-galactose-4-epimerase which catalyzes 2 distinct but analogous reactions: the epimerization of UDP-glucose to UDP-galactose, and the epimerization of UDP-N-acetylglucosamine to UDP-N-acetylgalactosamine. The bifunctional nature of the enzyme has the important metabolic consequence that mutant cells (or individuals) are dependent not only on exogenous galactose, but also on exogenous N-acetylgalactosamine for necessary precursor for the synthesis of glycoproteins and glycolipids. The missense mutations in the *GALE* gene result in the epimerase-deficiency galactosemia.

#### Breast Cancer and Sample Selection

[0040] Breast cancer is typically described as the uncontrolled growth of malignant breast tissue. Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory breast cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body.

[0041] As used herein, the term “breast cancer” refers to a condition characterized by anomalous rapid proliferation of abnormal cells in one or both breasts of a subject. The abnormal cells often are referred to as “neoplastic cells,” which are transformed cells that can form a solid tumor. The term “tumor” refers to an abnormal mass or population of cells (*i.e.* two or more cells) that result from excessive or abnormal cell division, whether malignant or benign, and pre-cancerous and cancerous cells. Malignant tumors are distinguished from benign growths or tumors in that, in addition to uncontrolled cellular proliferation, they can invade surrounding tissues and can metastasize. In breast cancer, neoplastic cells may be identified in one or both breasts only and not in another tissue or organ, in one or both breasts and one or more adjacent tissues or organs (*e.g.* lymph node), or in a breast and one or more non-adjacent tissues or organs to which the breast cancer cells have metastasized.

[0042] The term “invasion” as used herein refers to the spread of cancerous cells to adjacent surrounding tissues. The term “invasion” often is used synonymously with the term “metastasis,” which as used herein refers to a process in which cancer cells travel from one organ or tissue to another non-adjacent organ or tissue. Cancer cells in the breast(s) can spread to tissues and organs of a subject, and conversely, cancer cells from other organs or tissue can invade or metastasize to a breast. Cancerous cells from the breast(s) may invade or metastasize to any other organ or tissue of the body. Breast cancer

cells often invade lymph node cells and/or metastasize to the liver, brain and/or bone and spread cancer in these tissues and organs. Breast cancers can spread to other organs and tissues and cause lung cancer, prostate cancer, colon cancer, ovarian cancer, cervical cancer, gastrointestinal cancer, pancreatic cancer, glioblastoma, bladder cancer, hepatoma, colorectal cancer, uterine cervical cancer, endometrial carcinoma, salivary gland carcinoma, kidney cancer, vulval cancer, thyroid cancer, hepatic carcinoma, skin cancer, melanoma, ovarian cancer, neuroblastoma, myeloma, various types of head and neck cancer, acute lymphoblastic leukemia, acute myeloid leukemia, Ewing sarcoma and peripheral neuroepithelioma, and other carcinomas, lymphomas, blastomas, sarcomas, and leukemias.

[0043] Breast cancers arise most commonly in the lining of the milk ducts of the breast (ductal carcinoma), or in the lobules where breast milk is produced (lobular carcinoma). Other forms of breast cancer include Inflammatory Breast Cancer and Recurrent Breast Cancer. Inflammatory Breast Cancer is a rare, but very serious, aggressive type of breast cancer. The breast may look red and feel warm with ridges, welts, or hives on the breast; or the skin may look wrinkled. It is sometimes misdiagnosed as a simple infection. Recurrent disease means that the cancer has come back after it has been treated. It may come back in the breast, in the soft tissues of the chest (the chest wall), or in another part of the body. As used herein, the term “breast cancer” may include both Inflammatory Breast Cancer and Recurrent Breast Cancer.

[0044] In an effort to detect breast cancer as early as possible, regular physical exams and screening mammograms often are prescribed and conducted. A diagnostic mammogram often is performed to evaluate a breast complaint or abnormality detected by physical exam or routine screening mammography. If an abnormality seen with diagnostic mammography is suspicious, additional breast imaging (with exams such as ultrasound) or a biopsy may be ordered. A biopsy followed by pathological (microscopic) analysis is a definitive way to determine whether a subject has breast cancer. Excised breast cancer samples often are subjected to the following analyses: diagnosis of the breast tumor and confirmation of its malignancy; maximum tumor thickness; assessment of completeness of excision of invasive and *in situ* components and microscopic measurements of the shortest extent of clearance; level of invasion; presence and extent of regression; presence and extent of ulceration; histological type and special variants; pre-existing lesion; mitotic rate; vascular invasion; neurotropism; cell type; tumor lymphocyte infiltration; and growth phase.

[0045] The stage of a breast cancer can be classified as a range of stages from Stage 0 to Stage IV based on its size and the extent to which it has spread. The following table summarizes the stages:



**Table A**

<b>Stage</b>	<b>Tumor Size</b>	<b>Lymph Node Involvement</b>	<b>Metastasis (Spread)</b>
I	Less than 2 cm	No	No
II	Between 2-5 cm	No or in same side of breast	No
III	More than 5 cm	Yes, on same side of breast	No
IV	Not applicable	Not applicable	Yes

[0046] Stage 0 cancer is a contained cancer that has not spread beyond the breast ductal system. Fifteen to twenty percent of breast cancers detected by clinical examinations or testing are in Stage 0 (the earliest form of breast cancer). Two types of Stage 0 cancer are lobular carcinoma in situ (LCIS) and ductal carcinoma in situ (DCIS). LCIS indicates high risk for breast cancer. Many physicians do not classify LCIS as a malignancy and often encounter LCIS by chance on breast biopsy while investigating another area of concern. While the microscopic features of LCIS are abnormal and are similar to malignancy, LCIS does not behave as a cancer (and therefore is not treated as a cancer). LCIS is merely a marker for a significantly increased risk of cancer anywhere in the breast. However, bilateral simple mastectomy may be occasionally performed if LCIS patients have a strong family history of breast cancer. In DCIS the cancer cells are confined to milk ducts in the breast and have not spread into the fatty breast tissue or to any other part of the body (such as the lymph nodes). DCIS may be detected on mammogram as tiny specks of calcium (known as microcalcifications) 80% of the time. Less commonly DCIS can present itself as a mass with calcifications (15% of the time); and even less likely as a mass without calcifications (<5% of the time). A breast biopsy is used to confirm DCIS. A standard DCIS treatment is breast-conserving therapy (BCT), which is lumpectomy followed by radiation treatment or mastectomy. To date, DCIS patients have chosen equally among lumpectomy and mastectomy as their treatment option, though specific cases may sometimes favor lumpectomy over mastectomy or vice versa.

[0047] In Stage I, the primary (original) cancer is 2 cm or less in diameter and has not spread to the lymph nodes. In Stage IIA, the primary tumor is between 2 and 5 cm in diameter and has not spread to the lymph nodes. In Stage IIB, the primary tumor is between 2 and 5 cm in diameter and has spread to the axillary (underarm) lymph nodes; or the primary tumor is over 5 cm and has not spread to the lymph nodes. In Stage IIIA, the primary breast cancer of any kind that has spread to the axillary (underarm) lymph nodes and to axillary tissues. In Stage IIIB, the primary breast cancer is any size, has attached

itself to the chest wall, and has spread to the pectoral (chest) lymph nodes. In Stage IV, the primary cancer has spread out of the breast to other parts of the body (such as bone, lung, liver, brain). The treatment of Stage IV breast cancer focuses on extending survival time and relieving symptoms.

[0048] Based in part upon selection criteria set forth above, individuals having breast cancer can be selected for genetic studies. Also, individuals having no history of cancer or breast cancer often are selected for genetic studies. Other selection criteria can include: a tissue or fluid sample is derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, *e.g.*, lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy).

[0049] Provided herein is a set of blood samples and a set of corresponding nucleic acid samples isolated from the blood samples, where the blood samples are donated from individuals diagnosed with breast cancer. The sample set often includes blood samples or nucleic acid samples from 100 or more, 150 or more, or 200 or more individuals having breast cancer, and sometimes from 250 or more, 300 or more, 400 or more, or 500 or more individuals. The individuals can have parents from any place of origin, and in an embodiment, the set of samples are extracted from individuals of German paternal and German maternal ancestry. The samples in each set may be selected based upon five or more criteria and/or phenotypes set forth above.

#### Polymorphic Variants Associated with Breast Cancer

[0050] A genetic analysis provided herein linked breast cancer with polymorphic variants in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE* regions of the human genome disclosed herein. As used herein, the term “polymorphic site” refers to a region in a nucleic acid at which two or more alternative nucleotide sequences are observed in a significant number of nucleic acid samples from a population of individuals. A polymorphic site may be a nucleotide sequence of two or more nucleotides, an inserted nucleotide or nucleotide sequence, a deleted nucleotide or nucleotide sequence, or a microsatellite, for example. A polymorphic site that is two or more nucleotides in length may be 3, 4, 5,

6, 7, 8, 9, 10, 11, 12, 13, 14, 15 or more, 20 or more, 30 or more, 50 or more, 75 or more, 100 or more, 500 or more, or about 1000 nucleotides in length, where all or some of the nucleotide sequences differ within the region. A polymorphic site is often one nucleotide in length, which is referred to herein as a “single nucleotide polymorphism” or a “SNP.”

[0051] Where there are two, three, or four alternative nucleotide sequences at a polymorphic site, each nucleotide sequence is referred to as a “polymorphic variant” or “nucleic acid variant.” Where two polymorphic variants exist, for example, the polymorphic variant represented in a minority of samples from a population is sometimes referred to as a “minor allele” and the polymorphic variant that is more prevalently represented is sometimes referred to as a “major allele.” Many organisms possess a copy of each chromosome (*e.g.*, humans), and those individuals who possess two major alleles or two minor alleles are often referred to as being “homozygous” with respect to the polymorphism, and those individuals who possess one major allele and one minor allele are normally referred to as being “heterozygous” with respect to the polymorphism. Individuals who are homozygous with respect to one allele are sometimes predisposed to a different phenotype as compared to individuals who are heterozygous or homozygous with respect to another allele.

[0052] Furthermore, a genotype or polymorphic variant may be expressed in terms of a “haplotype,” which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0053] As used herein, the term “phenotype” refers to a trait which can be compared between individuals, such as presence or absence of a condition, a visually observable difference in appearance between individuals, metabolic variations, physiological variations, variations in the function of biological molecules, and the like. An example of a phenotype is occurrence of breast cancer.

[0054] Researchers sometimes report a polymorphic variant in a database without determining whether the variant is represented in a significant fraction of a population. Because a subset of these reported polymorphic variants are not represented in a statistically significant portion of the population, some of them are sequencing errors and/or not biologically relevant. Thus, it is often not known whether a reported polymorphic variant is statistically significant or biologically relevant until the presence of the variant is detected in a population of individuals and the frequency of the variant is determined. Methods for detecting a polymorphic variant in a population are described herein, specifically in Example 2. A

polymorphic variant is statistically significant and often biologically relevant if it is represented in 5% or more of a population, sometimes 10% or more, 15% or more, or 20% or more of a population, and often 25% or more, 30% or more, 35% or more, 40% or more, 45% or more, or 50% or more of a population.

[0055] A polymorphic variant may be detected on either or both strands of a double-stranded nucleic acid. For example, a thymine at a particular position in SEQ ID NO: 1 can be reported as an adenine from the complementary strand. Also, a polymorphic variant may be located within an intron or exon of a gene or within a portion of a regulatory region such as a promoter, a 5' untranslated region (UTR), a 3' UTR, and in DNA (*e.g.*, genomic DNA (gDNA) and complementary DNA (cDNA)), RNA (*e.g.*, mRNA, tRNA, and rRNA), or a polypeptide. Polymorphic variations may or may not result in detectable differences in gene expression, polypeptide structure, or polypeptide function.

[0056] In the genetic analysis that associated breast cancer with the polymorphic variants described hereafter, samples from individuals having breast cancer and individuals not having cancer were allelotyped and genotyped. The term "genotyped" as used herein refers to a process for determining a genotype of one or more individuals, where a "genotype" is a representation of one or more polymorphic variants in a population. Genotypes may be expressed in terms of a "haplotype," which as used herein refers to two or more polymorphic variants occurring within genomic DNA in a group of individuals within a population. For example, two SNPs may exist within a gene where each SNP position includes a cytosine variation and an adenine variation. Certain individuals in a population may carry one allele (heterozygous) or two alleles (homozygous) having the gene with a cytosine at each SNP position. As the two cytosines corresponding to each SNP in the gene travel together on one or both alleles in these individuals, the individuals can be characterized as having a cytosine/cytosine haplotype with respect to the two SNPs in the gene.

[0057] It was determined that polymorphic variations associated with an increased risk of breast cancer existed in *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequences. Polymorphic variants in and around the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE* loci were tested for association with breast cancer. In the *ICAM* locus, these included polymorphic variants at positions in SEQ ID NO: 1 selected from the group consisting of 139, 11799, 11851, 11851, 11963, 24282, 26849, 29633, 31254, 31967, 32920, 33929, 35599, 36101, 36101, 36340, 36405, 36517, 36777, 36992, 37645, 37868, 38440, 38440, 38532, 38532, 38547, 38547, 38712, 40684, 40860, 41213, 41419, 41613, 42407, 43440, 43440, 44247, 44247, 44247, 44247, 44677, 44677, 45256, 45256, 45536, 45536, 46153, 47546, 47697, 47944, 47944, 48530, 51102, 57090, 60093, 60439, 62694, 66260, 67295, 67295, 67304, 67731, 67731, 68555, 68555, 70429, 70875, 72360, 74228, 76802, 77664, 78803, 79263, 80810, 81020, 82426, 82783, 85912, 85912, 86135, 86135, 87877, 87877, 88043, 88043, 88206, 88343, 90701, 90701, 90974, 91060, 91087, 91594, 91594, 92302, 92384, 36517, and 44677. Polymorphic variants in a region

spanning positions 11851-24282, 36340-37868, 41213-41613, 70875-74228, 42407-45536, and 42407-51102 in SEQ ID NO: 1 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 11963, 36340, 36992, 37868, 41213, 41419, 41613, 42407, 44247, 44677, 45256, 45536, 51102, 72360, 36517, and 44677 in SEQ ID NO: 1. At these positions in SEQ ID NO: 1, an adenine at position 11963, a guanine at position 36340, an adenine at position 36992, a guanine at position 37868, a cytosine at position 41213, a guanine at position 41419, a guanine at position 41613, a cytosine at position 42407, a cytosine at position 44247, an adenine or cytosine at position 44677, a thymine at position 45256, a guanine at position 45536, a cytosine at position 51102, a guanine at position 72360, a cytosine at position 36517, and guanine at position 44677, in particular were associated with risk of breast cancer. Also, a proline at amino acid position 352 or an alanine at amino acid position 348 in SEQ ID NO: 15 were in particular associated with an increased risk of breast cancer.

[0058] In the *MAPK10* locus, these included polymorphic variants at positions in SEQ ID NO: 2 selected from the group consisting of 191, 1490, 3781, 3935, 4512, 7573, 8467, 9001, 9732, 13477, 13787, 13903, 14355, 15053, 15459, 17762, 19482, 19631, 22170, 22688, 22748, 23376, 23826, 23868, 24154, 25972, 26057, 26361, 26599, 26712, 26812, 27069, 32421, 33557, 35127, 35222, 35999, 36424, 37403, 39203, 39226, 41147, 46176, 50452, 52919, 60214, 61093, 62572, 63601, 65362, 65863, 66207, 66339, 69512, 70759, 71217, 73382, and 76307. Polymorphic variants in a region spanning positions 23826-36424, 46176-62572, 4512-8467 or 13787-14355 in SEQ ID NO: 2 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 7573, 13903, 23826, 26057, 26361, 26599, 26812, 27069, 35127, 35222, 36424, 46176, 50452, 61093, 62572, and 70759 in SEQ ID NO: 2. At these positions in SEQ ID NO: 2, a guanine at position 7573, a cytosine at position 13903, an adenine at position 23826, an adenine at position 26057, a thymine at position 26361, an adenine at position 26599, an adenine at position 26812, a cytosine at position 27069, an adenine at position 35127, a thymine at position 35222, a cytosine at position 36424, a cytosine at position 46176, a cytosine at position 50452, a guanine at position 61093, an adenine at position 62572, and a guanine at position 70759, in particular were associated with risk of breast cancer.

[0059] In the *KIAA0861* locus, these included polymorphic variants at positions in SEQ ID NO: 3 selected from the group consisting of 107, 2157, 7300, 8233, 9647, 9868, 9889, 10621, 11003, 11507, 11527, 11718, 11808, 12024, 13963, 14300, 14361, 16287, 18635, 19365, 24953, 25435, 26847, 27492, 27620, 27678, 27714, 29719, 30234, 31909, 32153, 33572, 42164, 43925, 45031, 45655, 48350, 48418, 48563, 53189, 56468, 59358, 63761, 65931, 67040, 69491, 83308, 126545, 137592, and 147169. Polymorphic variants in a region spanning positions 42164-48563 in SEQ ID NO: 3 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 107, 42164, 45031, 45655, 48563, 19365 and 14361 in SEQ ID NO: 3. At these positions in SEQ ID NO: 3,

an adenine at position 107, a thymine at position 14361, a guanine at position 19365, a thymine at position 42164, a cytosine at position 45031, a thymine at position 45655 and a cytosine at position 48563, in particular were associated with risk of breast cancer. Also, leucine at amino acid position 359 in SEQ ID NO: 17, a leucine at amino acid position 378 in SEQ ID NO: 17, or an alanine at amino acid position 857 in SEQ ID NO: 17 were in particular associated with an increased risk of breast cancer.

[0060] In the *NUMA1* locus, these included polymorphic variants at positions in SEQ ID NO: 4 selected from the group consisting of 174, 815, 3480, 9715, 14755, 15912, 19834, 19850, 20171, 20500, 20536, 23187, 25289, 25470, 28720, 29566, 30155, 30752, 32710, 32954, 33725, 33842, 36345, 38115, 39150, 40840, 41969, 42045, 43785, 44444, 44579, 45386, 46827, 47320, 47625, 47837, 47866, 49002, 49566, 52058, 52249, 52257, 52850, 53860, 54052, 54411, 55098, 55303, 59398, 59533, 60542, 61541, 62309, 72299, 73031, 73803, 80950, 82137, 96077, 96470, 98116, 98184, and 132952. Polymorphic variants in a region spanning positions 174-32954, 38115-43785, 45386-52058, 52257-54411, 55303-73803 or 96470-98184 in SEQ ID NO: 4 in particular were associated with an increased risk of breast cancer, including polymorphic variants at positions 174, 815, 3480, 19834, 19850, 20171, 20500, 20536, 23187, 25470, 30155, 30752, 32710, 32954, 38115, 39150, 40840, 41969, 42045, 43785, 45386, 46827, 47320, 47625, 47837, 47866, 49002, 49566, 52058, 52257, 52850, 53860, 54052, 54411, 55303, 59398, 60542, 62309, 72299, 73031, 73803, and 98116 in SEQ ID NO: 4. At these positions in SEQ ID NO: 4, a thymine at position 174, an adenine at position 815, a cytosine at position 3480, a guanine at position 19834, an adenine at position 19850, a thymine at position 20171, a thymine at position 20500, a cytosine at position 20536, a cytosine at position 23187, a thymine at position 25470, a thymine at position 30155, a guanine at position 30752, a thymine at position 32710, a guanine at position 32954, an adenine at position 38115, a cytosine at position 39150, a thymine at position 40840, an adenine at position 41969, a thymine at position 42045, a guanine at position 43785, a cytosine at position 45386, an adenine at position 46827, an adenine at position 47320, a cytosine at position 47625, a cytosine at position 47837, an adenine at position 47866, a cytosine at position 49002, a thymine at position 49566, a cytosine at position 52058, a thymine at position 52257, a thymine at position 52850, a cytosine at position 53860, a cytosine at position 54052, a thymine at position 54411, a cytosine at position 55303, an adenine at position 59398, an adenine at position 60542, an adenine at position 62309, a cytosine at position 72299, a thymine at position 73031, a guanine at position 73803, and a thymine at position 98116, in particular were associated with risk of breast cancer. In the *GALE* locus, a polymorphic variant at position 174 in SEQ ID NO: 5 was in particular associated with increased risk of breast cancer, and an adenine this position was the cancer-associated allele.

Additional Polymorphic Variants Associated with Breast Cancer

[0061] Also provided is a method for identifying polymorphic variants proximal to an incident, founder polymorphic variant associated with breast cancer. Thus, featured herein are methods for identifying a polymorphic variation associated with breast cancer that is proximal to an incident polymorphic variation associated with breast cancer, which comprises identifying a polymorphic variant proximal to the incident polymorphic variant associated with breast cancer, where the incident polymorphic variant is in a nucleotide sequence set forth in SEQ ID NO: 1-5. The nucleotide sequence often comprises a polynucleotide sequence selected from the group consisting of (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), often a fragment that includes a polymorphic site associated with breast cancer. The presence or absence of an association of the proximal polymorphic variant with breast cancer then is determined using a known association method, such as a method described in the Examples hereafter. In an embodiment, the incident polymorphic variant is described in SEQ ID NO: 1-5. In another embodiment, the proximal polymorphic variant identified sometimes is a publicly disclosed polymorphic variant, which for example, sometimes is published in a publicly available database. In other embodiments, the polymorphic variant identified is not publicly disclosed and is discovered using a known method, including, but not limited to, sequencing a region surrounding the incident polymorphic variant in a group of nucleic acid samples. Thus, multiple polymorphic variants proximal to an incident polymorphic variant are associated with breast cancer using this method.

[0062] The proximal polymorphic variant often is identified in a region surrounding the incident polymorphic variant. In certain embodiments, this surrounding region is about 50 kb flanking the first polymorphic variant (*e.g.* about 50 kb 5' of the first polymorphic variant and about 50 kb 3' of the first polymorphic variant), and the region sometimes is composed of shorter flanking sequences, such as flanking sequences of about 40 kb, about 30 kb, about 25 kb, about 20 kb, about 15 kb, about 10 kb, about 7 kb, about 5 kb, or about 2 kb 5' and 3' of the incident polymorphic variant. In other embodiments, the region is composed of longer flanking sequences, such as flanking sequences of about 55 kb, about 60 kb, about 65 kb, about 70 kb, about 75 kb, about 80 kb, about 85 kb, about 90 kb, about 95 kb, or about 100 kb 5' and 3' of the incident polymorphic variant.

[0063] In certain embodiments, polymorphic variants associated with breast cancer are identified iteratively. For example, a first proximal polymorphic variant is associated with breast cancer using the

methods described above and then another polymorphic variant proximal to the first proximal polymorphic variant is identified (*e.g.*, publicly disclosed or discovered) and the presence or absence of an association of one or more other polymorphic variants proximal to the first proximal polymorphic variant with breast cancer is determined.

[0064] The methods described herein are useful for identifying or discovering additional polymorphic variants that may be used to further characterize a gene, region or loci associated with a condition, a disease (*e.g.*, breast cancer), or a disorder. For example, allelotyping or genotyping data from the additional polymorphic variants may be used to identify a functional mutation or a region of linkage disequilibrium.

[0065] In certain embodiments, polymorphic variants identified or discovered within a region comprising the first polymorphic variant associated with breast cancer are genotyped using the genetic methods and sample selection techniques described herein, and it can be determined whether those polymorphic variants are in linkage disequilibrium with the first polymorphic variant. The size of the region in linkage disequilibrium with the first polymorphic variant also can be assessed using these genotyping methods. Thus, provided herein are methods for determining whether a polymorphic variant is in linkage disequilibrium with a first polymorphic variant associated with breast cancer, and such information can be used in prognosis methods described herein.

Isolated *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* Nucleic Acids

[0066] Featured herein are isolated *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acids, which include the nucleic acid having the nucleotide sequence of SEQ ID NO: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 or 11, nucleic acid variants, and substantially identical nucleic acids of the foregoing. Nucleotide sequences of the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acids sometimes are referred to herein as “*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequences.” A “*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid variant” refers to one allele that may have one or more different polymorphic variations as compared to another allele in another subject or the same subject. A polymorphic variation in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid variant may be represented on one or both strands in a double-stranded nucleic acid or on one chromosomal complement (heterozygous) or both chromosomal complements (homozygous).

[0067] As used herein, the term “nucleic acid” includes DNA molecules (*e.g.*, a complementary DNA (cDNA) and genomic DNA (gDNA)) and RNA molecules (*e.g.*, mRNA, rRNA, and tRNA) and analogs of DNA or RNA, for example, by use of nucleotide analogs. The nucleic acid molecule can be single-stranded and it is often double-stranded. The term “isolated or purified nucleic acid” refers to nucleic acids that are separated from other nucleic acids present in the natural source of the nucleic acid.



For example, with regard to genomic DNA, the term “isolated” includes nucleic acids which are separated from the chromosome with which the genomic DNA is naturally associated. An “isolated” nucleic acid is often free of sequences which naturally flank the nucleic acid (i.e., sequences located at the 5’ and/or 3’ ends of the nucleic acid) in the genomic DNA of the organism from which the nucleic acid is derived. For example, in various embodiments, the isolated nucleic acid molecule can contain less than about 5 kb, 4 kb, 3 kb, 2 kb, 1 kb, 0.5 kb or 0.1 kb of 5’ and/or 3’ nucleotide sequences which flank the nucleic acid molecule in genomic DNA of the cell from which the nucleic acid is derived. Moreover, an “isolated” nucleic acid molecule, such as a cDNA molecule, can be substantially free of other cellular material, or culture medium when produced by recombinant techniques, or substantially free of chemical precursors or other chemicals when chemically synthesized. As used herein, the term “*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene” refers to a nucleotide sequence that encodes a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide.

[0068] Also included herein are nucleic acid fragments. These fragments typically are a nucleotide sequence identical to a nucleotide sequence in SEQ ID NO: 1-12, a nucleotide sequence substantially identical to a nucleotide sequence in SEQ ID NO: 1-12, or a nucleotide sequence that is complementary to the foregoing. The nucleic acid fragment may be identical, substantially identical or homologous to a nucleotide sequence in an exon or an intron in SEQ ID NO: 1-5, and may encode a domain or part of a domain or motif of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Sometimes, the fragment will comprises the polymorphic variation described herein as being associated with breast cancer. The nucleic acid fragment sometimes is 50, 100, or 200 or fewer base pairs in length, and is sometimes about 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900, 2000, 2100, 2200, 2300, 2400, 2500, 2600, 2700, 2800, 2900, 3000, 3100, 3200, 3300, 3400, 3500, 3600, 3800, 4000, 5000, 6000, 7000, 8000, 9000, 10000, 15000, 20000, 30000, 40000, 50000, 60000, 70000, 80000, 90000, 100000, 110000, 120000, 130000, 140000, 150000 or 160000 base pairs in length. A nucleic acid fragment complementary to a nucleotide sequence identical or substantially identical to the nucleotide sequence of SEQ ID NO: 1-12 and hybridizes to such a nucleotide sequence under stringent conditions often is referred to as a “probe.” Nucleic acid fragments often include one or more polymorphic sites, or sometimes have an end that is adjacent to a polymorphic site as described hereafter.

[0069] An example of a nucleic acid fragment is an oligonucleotide. As used herein, the term “oligonucleotide” refers to a nucleic acid comprising about 8 to about 50 covalently linked nucleotides, often comprising from about 8 to about 35 nucleotides, and more often from about 10 to about 25 nucleotides. The backbone and nucleotides within an oligonucleotide may be the same as those of naturally occurring nucleic acids, or analogs or derivatives of naturally occurring nucleic acids, provided

that oligonucleotides having such analogs or derivatives retain the ability to hybridize specifically to a nucleic acid comprising a targeted polymorphism. Oligonucleotides described herein may be used as hybridization probes or as components of prognostic or diagnostic assays, for example, as described herein.

[0070] Oligonucleotides are typically synthesized using standard methods and equipment, such as the ABI 3900 High Throughput DNA Synthesizer and the EXPEDITE™ 8909 Nucleic Acid Synthesizer, both of which are available from Applied Biosystems (Foster City, CA). Analogs and derivatives are exemplified in U.S. Pat. Nos. 4,469,863; 5,536,821; 5,541,306; 5,637,683; 5,637,684; 5,700,922; 5,717,083; 5,719,262; 5,739,308; 5,773,601; 5,886,165; 5,929,226; 5,977,296; 6,140,482; WO 00/56746; WO 01/14398, and related publications. Methods for synthesizing oligonucleotides comprising such analogs or derivatives are disclosed, for example, in the patent publications cited above and in U.S. Pat. Nos. 5,614,622; 5,739,314; 5,955,599; 5,962,674; 6,117,992; in WO 00/75372; and in related publications.

[0071] Oligonucleotides also may be linked to a second moiety. The second moiety may be an additional nucleotide sequence such as a tail sequence (e.g., a polyadenosine tail), an adapter sequence (e.g., phage M13 universal tail sequence), and others. Alternatively, the second moiety may be a non-nucleotide moiety such as a moiety which facilitates linkage to a solid support or a label to facilitate detection of the oligonucleotide. Such labels include, without limitation, a radioactive label, a fluorescent label, a chemiluminescent label, a paramagnetic label, and the like. The second moiety may be attached to any position of the oligonucleotide, provided the oligonucleotide can hybridize to the nucleic acid comprising the polymorphism.

#### Uses for Nucleic Acid Sequences

[0072] Nucleic acid coding sequences depicted in SEQ ID NO: 1-12 may be used for diagnostic purposes for detection and control of polypeptide expression. Also, included herein are oligonucleotide sequences such as antisense RNA, small-interfering RNA (siRNA) and DNA molecules and ribozymes that function to inhibit translation of a polypeptide. Antisense techniques and RNA interference techniques are known in the art and are described herein.

[0073] Ribozymes are enzymatic RNA molecules capable of catalyzing the specific cleavage of RNA. The mechanism of ribozyme action involves sequence specific hybridization of the ribozyme molecule to complementary target RNA, followed by an endonucleolytic cleavage. Ribozymes may be engineered hammerhead motif ribozyme molecules that specifically and efficiently catalyze endonucleolytic cleavage of RNA sequences corresponding to or complementary to the nucleotide sequences set forth in SEQ ID NO: 1-12. Specific ribozyme cleavage sites within any potential RNA

target are initially identified by scanning the target molecule for ribozyme cleavage sites which include the following sequences, GUA, GUU and GUC. Once identified, short RNA sequences of between fifteen (15) and twenty (20) ribonucleotides corresponding to the region of the target gene containing the cleavage site may be evaluated for predicted structural features such as secondary structure that may render the oligonucleotide sequence unsuitable. The suitability of candidate targets may also be evaluated by testing their accessibility to hybridization with complementary oligonucleotides, using ribonuclease protection assays.

[0074] Antisense RNA and DNA molecules, siRNA and ribozymes may be prepared by any method known in the art for the synthesis of RNA molecules. These include techniques for chemically synthesizing oligodeoxyribonucleotides well known in the art such as solid phase phosphoramidite chemical synthesis. Alternatively, RNA molecules may be generated by *in vitro* and *in vivo* transcription of DNA sequences encoding the antisense RNA molecule. Such DNA sequences may be incorporated into a wide variety of vectors which incorporate suitable RNA polymerase promoters such as the T7 or SP6 polymerase promoters. Alternatively, antisense cDNA constructs that synthesize antisense RNA constitutively or inducibly, depending on the promoter used, can be introduced stably into cell lines.

[0075] DNA encoding a polypeptide also may have a number of uses for the diagnosis of diseases, including breast cancer, resulting from aberrant expression of a target gene described herein. For example, the nucleic acid sequence may be used in hybridization assays of biopsies or autopsies to diagnose abnormalities of expression or function (*e.g.*, Southern or Northern blot analysis, *in situ* hybridization assays).

[0076] In addition, the expression of a polypeptide during embryonic development may also be determined using nucleic acid encoding the polypeptide. As addressed, *infra*, production of functionally impaired polypeptide can be the cause of various disease states, such as breast cancer. *In situ* hybridizations using polynucleotide probes may be employed to predict problems related to breast cancer. Further, as indicated, *infra*, administration of human active polypeptide, recombinantly produced as described herein, may be used to treat disease states related to functionally impaired polypeptide. Alternatively, gene therapy approaches may be employed to remedy deficiencies of functional polypeptide or to replace or compete with dysfunctional polypeptide.

#### Expression Vectors, Host Cells, and Genetically Engineered Cells

[0077] Provided herein are nucleic acid vectors, often expression vectors, which contain a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid. As used herein, the term “vector” refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked and can include a plasmid, cosmid, or viral vector. The vector can be capable of autonomous replication or it can

integrate into a host DNA. Viral vectors may include replication defective retroviruses, adenoviruses and adeno-associated viruses for example.

[0078] A vector can include a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid in a form suitable for expression of the nucleic acid in a host cell. The recombinant expression vector typically includes one or more regulatory sequences operatively linked to the nucleic acid sequence to be expressed. The term "regulatory sequence" includes promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Regulatory sequences include those that direct constitutive expression of a nucleotide sequence, as well as tissue-specific regulatory and/or inducible sequences. The design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of polypeptide desired, and the like. Expression vectors can be introduced into host cells to produce *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides, including fusion polypeptides, encoded by *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acids.

[0079] Recombinant expression vectors can be designed for expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides in prokaryotic or eukaryotic cells. For example, *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides can be expressed in *E. coli*, insect cells (e.g., using baculovirus expression vectors), yeast cells, or mammalian cells. Suitable host cells are discussed further in Goeddel, *Gene Expression Technology: Methods in Enzymology* 185, Academic Press, San Diego, CA (1990). Alternatively, the recombinant expression vector can be transcribed and translated in vitro, for example using T7 promoter regulatory sequences and T7 polymerase.

[0080] Expression of polypeptides in prokaryotes is most often carried out in *E. coli* with vectors containing constitutive or inducible promoters directing the expression of either fusion or non-fusion polypeptides. Fusion vectors add a number of amino acids to a polypeptide encoded therein, usually to the amino terminus of the recombinant polypeptide. Such fusion vectors typically serve three purposes: 1) to increase expression of recombinant polypeptide; 2) to increase the solubility of the recombinant polypeptide; and 3) to aid in the purification of the recombinant polypeptide by acting as a ligand in affinity purification. Often, a proteolytic cleavage site is introduced at the junction of the fusion moiety and the recombinant polypeptide to enable separation of the recombinant polypeptide from the fusion moiety subsequent to purification of the fusion polypeptide. Such enzymes, and their cognate recognition sequences, include Factor Xa, thrombin and enterokinase. Typical fusion expression vectors include pGEX (Pharmacia Biotech Inc; Smith & Johnson, *Gene* 67: 31-40 (1988)), pMAL (New England Biolabs, Beverly, MA) and pRIT5 (Pharmacia, Piscataway, NJ) which fuse glutathione S-transferase (GST), maltose E binding polypeptide, or polypeptide A, respectively, to the target recombinant polypeptide.

[0081] Purified fusion polypeptides can be used in screening assays and to generate antibodies specific for *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides. In a therapeutic embodiment, fusion polypeptide expressed in a retroviral expression vector is used to infect bone marrow cells that are subsequently transplanted into irradiated recipients. The pathology of the subject recipient is then examined after sufficient time has passed (e.g., six (6) weeks).

[0082] Expressing the polypeptide in host bacteria with an impaired capacity to proteolytically cleave the recombinant polypeptide is often used to maximize recombinant polypeptide expression (Gottesman, S., *Gene Expression Technology: Methods in Enzymology*, Academic Press, San Diego, California 185: 119-128 (1990)). Another strategy is to alter the nucleotide sequence of the nucleic acid to be inserted into an expression vector so that the individual codons for each amino acid are those preferentially utilized in *E. coli* (Wada et al., *Nucleic Acids Res.* 20: 2111-2118 (1992)). Such alteration of nucleotide sequences can be carried out by standard DNA synthesis techniques.

[0083] When used in mammalian cells, the expression vector's control functions are often provided by viral regulatory elements. For example, commonly used promoters are derived from polyoma, Adenovirus 2, cytomegalovirus and Simian Virus 40. Recombinant mammalian expression vectors are often capable of directing expression of the nucleic acid in a particular cell type (e.g., tissue-specific regulatory elements are used to express the nucleic acid). Non-limiting examples of suitable tissue-specific promoters include an albumin promoter (liver-specific; Pinkert et al., *Genes Dev.* 1: 268-277 (1987)), lymphoid-specific promoters (Calame & Eaton, *Adv. Immunol.* 43: 235-275 (1988)), promoters of T cell receptors (Winoto & Baltimore, *EMBO J.* 8: 729-733 (1989)) promoters of immunoglobulins (Banerji et al., *Cell* 33: 729-740 (1983); Queen & Baltimore, *Cell* 33: 741-748 (1983)), neuron-specific promoters (e.g., the neurofilament promoter; Byrne & Ruddle, *Proc. Natl. Acad. Sci. USA* 86: 5473-5477 (1989)), pancreas-specific promoters (Edlund et al., *Science* 230: 912-916 (1985)), and mammary gland-specific promoters (e.g., milk whey promoter; U.S. Patent No. 4,873,316 and European Application Publication No. 264,166). Developmentally-regulated promoters are sometimes utilized, for example, the murine hox promoters (Kessel & Gruss, *Science* 249: 374-379 (1990)) and the  $\alpha$ -fetopolypeptide promoter (Campes & Tilghman, *Genes Dev.* 3: 537-546 (1989)).

[0084] A *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid may also be cloned into an expression vector in an antisense orientation. Regulatory sequences (e.g., viral promoters and/or enhancers) operatively linked to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid cloned in the antisense orientation can be chosen for directing constitutive, tissue specific or cell type specific expression of antisense RNA in a variety of cell types. Antisense expression vectors can be in the form of a recombinant plasmid, phagemid or attenuated virus. For a discussion of the regulation of gene

expression using antisense genes see Weintraub et al., Antisense RNA as a molecular tool for genetic analysis, Reviews - Trends in Genetics, Vol. 1(1) (1986).

[0085] Also provided herein are host cells that include a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid within a recombinant expression vector or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid sequence fragments which allow it to homologously recombine into a specific site of the host cell genome. The terms “host cell” and “recombinant host cell” are used interchangeably herein. Such terms refer not only to the particular subject cell but rather also to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein. A host cell can be any prokaryotic or eukaryotic cell. For example, a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide can be expressed in bacterial cells such as *E. coli*, insect cells, yeast or mammalian cells (such as Chinese hamster ovary cells (CHO) or COS cells). Other suitable host cells are known to those skilled in the art.

[0086] Vectors can be introduced into host cells via conventional transformation or transfection techniques. As used herein, the terms “transformation” and “transfection” are intended to refer to a variety of art-recognized techniques for introducing foreign nucleic acid (e.g., DNA) into a host cell, including calcium phosphate or calcium chloride co-precipitation, transduction/infection, DEAE-dextran-mediated transfection, lipofection, or electroporation.

[0087] A host cell provided herein can be used to produce (i.e., express) a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Accordingly, further provided are methods for producing a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide using the host cells described herein. In one embodiment, the method includes culturing host cells into which a recombinant expression vector encoding a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide has been introduced in a suitable medium such that a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide is produced. In another embodiment, the method further includes isolating a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide from the medium or the host cell.

[0088] Also provided are cells or purified preparations of cells which include a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene, or which otherwise misexpress *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Cell preparations can consist of human or non-human cells, e.g., rodent cells, e.g., mouse or rat cells, rabbit cells, or pig cells. In certain embodiments, the cell or cells include a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene (e.g., a heterologous form of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* such as a human gene expressed in non-human cells). The *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene can be misexpressed, e.g., overexpressed or underexpressed. In other embodiments, the cell or cells include a gene which misexpress an endogenous

*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide (e.g., expression of a gene is disrupted, also known as a knockout). Such cells can serve as a model for studying disorders which are related to mutated or mis-expressed *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* alleles or for use in drug screening. Also provided are human cells (e.g., a hematopoietic stem cells) transformed with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid.

[0089] Also provided are cells or a purified preparation thereof (e.g., human cells) in which an endogenous *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid is under the control of a regulatory sequence that does not normally control the expression of the endogenous *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene. The expression characteristics of an endogenous gene within a cell (e.g., a cell line or microorganism) can be modified by inserting a heterologous DNA regulatory element into the genome of the cell such that the inserted regulatory element is operably linked to the endogenous *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene. For example, an endogenous *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene (e.g., a gene which is “transcriptionally silent,” not normally expressed, or expressed only at very low levels) may be activated by inserting a regulatory element which is capable of promoting the expression of a normally expressed gene product in that cell. Techniques such as targeted homologous recombinations, can be used to insert the heterologous DNA as described in, e.g., Chappel, US 5,272,071; WO 91/06667, published on May 16, 1991.

#### Transgenic Animals

[0090] Non-human transgenic animals that express a heterologous *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide (e.g., expressed from a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid isolated from another organism) can be generated. Such animals are useful for studying the function and/or activity of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide and for identifying and/or evaluating modulators of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid and *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide activity. As used herein, a “transgenic animal” is a non-human animal such as a mammal (e.g., a non-human primate such as chimpanzee, baboon, or macaque; an ungulate such as an equine, bovine, or caprine; or a rodent such as a rat, a mouse, or an Israeli sand rat), a bird (e.g., a chicken or a turkey), an amphibian (e.g., a frog, salamander, or newt), or an insect (e.g., *Drosophila melanogaster*), in which one or more of the cells of the animal includes a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene. A transgene is exogenous DNA or a rearrangement (e.g., a deletion of endogenous chromosomal DNA) that is often integrated into or occurs in the genome of cells in a transgenic animal. A transgene can direct expression of an encoded gene product in one or more cell types or tissues of the transgenic animal, and other transgenes can reduce expression (e.g., a knockout). Thus, a transgenic animal can be one in which an endogenous

*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene has been altered by homologous recombination between the endogenous gene and an exogenous DNA molecule introduced into a cell of the animal (e.g., an embryonic cell of the animal) prior to development of the animal.

[0091] Intronic sequences and polyadenylation signals can also be included in the transgene to increase expression efficiency of the transgene. One or more tissue-specific regulatory sequences can be operably linked to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene to direct expression of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide to particular cells. A transgenic founder animal can be identified based upon the presence of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* transgene in its genome and/or expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA in tissues or cells of the animals. A transgenic founder animal can then be used to breed additional animals carrying the transgene. Moreover, transgenic animals carrying a transgene encoding a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide can further be bred to other transgenic animals carrying other transgenes.

[0092] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides can be expressed in transgenic animals or plants by introducing, for example, a nucleic acid encoding the polypeptide into the genome of an animal. In certain embodiments the nucleic acid is placed under the control of a tissue specific promoter, e.g., a milk or egg specific promoter, and recovered from the milk or eggs produced by the animal. Also included is a population of cells from a transgenic animal.

#### *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE* Polypeptides

[0093] Featured herein are isolated *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides, which include polypeptides having amino acid sequences set forth in SEQ ID NO: 13-18, and substantially identical polypeptides thereof. Such polypeptides sometimes are proteins or peptides. A *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide is a polypeptide encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid, where one nucleic acid can encode one or more different polypeptides. An “isolated” or “purified” polypeptide or protein is substantially free of cellular material or other contaminating proteins from the cell or tissue source from which the protein is derived, or substantially free from chemical precursors or other chemicals when chemically synthesized. In one embodiment, the language “substantially free” means preparation of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide variant having less than about 30%, 20%, 10% and sometimes 5% (by dry weight), of non-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide (also referred to herein as a “contaminating protein”), or of chemical precursors or non-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* chemicals. When the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or a biologically active portion thereof is



recombinantly produced, it is also often substantially free of culture medium, specifically, where culture medium represents less than about 20%, sometimes less than about 10%, and often less than about 5% of the volume of the polypeptide preparation. Isolated or purified *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide preparations are sometimes 0.01 milligrams or more or 0.1 milligrams or more, and often 1.0 milligrams or more and 10 milligrams or more in dry weight. In specific embodiments, a polypeptide comprises a leucine at amino acid position 359 in SEQ ID NO: 17, a leucine at amino acid position 378 in SEQ ID NO: 17, or an alanine at amino acid position 857 in SEQ ID NO: 17, or a *ICAM5* polypeptide comprises a proline at amino acid position 352 or an alanine at amino acid position 348 in SEQ ID NO: 15.

[0094] In another aspect, featured herein are *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides and biologically active or antigenic fragments thereof that are useful as reagents or targets in assays applicable to prevention, treatment or diagnosis of breast cancer. In another embodiment, provided herein are *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides having a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity or activities.

[0095] Further included herein are *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide fragments. The polypeptide fragment may be a domain or part of a domain of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. The polypeptide fragment is often 50 or fewer, 100 or fewer, or 200 or fewer amino acids in length, and is sometimes 300, 400, 500, 600, 700, or 900 or fewer amino acids in length. In certain embodiments, the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 1211 consecutive amino acids of SEQ ID NO: 13-18, or the polypeptide fragment comprises, consists essentially of, or consists of, at least 6 consecutive amino acids and not more than 543 consecutive amino acids of SEQ ID NO: 13-18.

[0096] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides described herein can be used as immunogens to produce anti-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antibodies in a subject, to purify *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* substrate. Full-length *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides and polynucleotides encoding the same may be specifically substituted for a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide fragment or polynucleotide encoding the same in any embodiment described herein.

[0097] Substantially identical polypeptides may depart from the amino acid sequences set forth in SEQ ID NO: 13-18 in different manners. For example, conservative amino acid modifications may be introduced at one or more positions in the amino acid sequences of SEQ ID NO: 13-18. A “conservative amino acid substitution” is one in which the amino acid is replaced by another amino acid having a

similar structure and/or chemical function. Families of amino acid residues having similar structures and functions are well known. These families include amino acids with basic side chains (e.g., lysine, arginine, histidine), acidic side chains (e.g., aspartic acid, glutamic acid), uncharged polar side chains (e.g., glycine, asparagine, glutamine, serine, threonine, tyrosine, cysteine), nonpolar side chains (e.g., alanine, valine, leucine, isoleucine, proline, phenylalanine, methionine, tryptophan), beta-branched side chains (e.g., threonine, valine, isoleucine) and aromatic side chains (e.g., tyrosine, phenylalanine, tryptophan, histidine). Also, essential and non-essential amino acids may be replaced. A “non-essential” amino acid is one that can be altered without abolishing or substantially altering the biological function of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, whereas altering an “essential” amino acid abolishes or substantially alters the biological function of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Amino acids that are conserved among *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides are typically essential amino acids.

[0098] Also, *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides and polypeptide variants may exist as chimeric or fusion polypeptides. As used herein, a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* “chimeric polypeptide” or “fusion polypeptide” includes a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide linked to a non-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. A “non-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide” refers to a polypeptide having an amino acid sequence corresponding to a polypeptide which is not substantially identical to the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, which includes, for example, a polypeptide that is different from the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide and derived from the same or a different organism. The *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide in the fusion polypeptide can correspond to an entire or nearly entire *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or a fragment thereof. The non-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide can be fused to the N-terminus or C-terminus of the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide.

[0099] Fusion polypeptides can include a moiety having high affinity for a ligand. For example, the fusion polypeptide can be a GST-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* fusion polypeptide in which the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* sequences are fused to the C-terminus of the GST sequences, or a polyhistidine-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* fusion polypeptide in which the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide is fused at the N- or C-terminus to a string of histidine residues. Such fusion polypeptides can facilitate purification of recombinant *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*. Expression vectors are commercially available that already encode a fusion moiety (e.g., a GST polypeptide), and a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid can be cloned into an expression vector such that the fusion moiety is linked in-frame to the

*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Further, the fusion polypeptide can be a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide containing a heterologous signal sequence at its N-terminus. In certain host cells (e.g., mammalian host cells), expression, secretion, cellular internalization, and cellular localization of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide can be increased through use of a heterologous signal sequence. Fusion polypeptides can also include all or a part of a serum polypeptide (e.g., an IgG constant region or human serum albumin).

[0100] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides or fragments thereof can be incorporated into pharmaceutical compositions and administered to a subject in vivo. Administration of these *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides can be used to affect the bioavailability of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* substrate and may effectively increase or decrease *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* biological activity in a cell or effectively supplement dysfunctional or hyperactive *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* fusion polypeptides may be useful therapeutically for the treatment of disorders caused by, for example, (i) aberrant modification or mutation of a gene encoding a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide; (ii) mis-regulation of the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* gene; and (iii) aberrant post-translational modification of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Also, *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides can be used as immunogens to produce anti-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antibodies in a subject, to purify *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* ligands or binding partners, and in screening assays to identify molecules which inhibit or enhance the interaction of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* substrate.

[0101] In addition, polypeptides can be chemically synthesized using techniques known in the art (See, e.g., Creighton, 1983 *Proteins*. New York, N.Y.: W. H. Freeman and Company; and Hunkapiller *et al.*, (1984) *Nature* July 12 -18;310(5973):105-11). For example, a relative short polypeptide fragment can be synthesized by use of a peptide synthesizer. Furthermore, if desired, non-classical amino acids or chemical amino acid analogs can be introduced as a substitution or addition into the fragment sequence. Non-classical amino acids include, but are not limited to, to the D-isomers of the common amino acids, 2,4-diaminobutyric acid, α-amino isobutyric acid, 4-aminobutyric acid, Abu, 2-amino butyric acid, g-Abu, e-Ahx, 6-amino hexanoic acid, Aib, 2-amino isobutyric acid, 3-amino propionic acid, ornithine, norleucine, norvaline, hydroxyproline, sarcosine, citrulline, homocitrulline, cysteic acid, t-butylglycine, t-butylalanine, phenylglycine, cyclohexylalanine, b-alanine, fluoroamino acids, designer amino acids such as b-methyl amino acids, Ca-methyl amino acids, Na-methyl amino acids, and amino acid analogs in general. Furthermore, the amino acid can be D (dextrorotary) or L (levorotary).

[0102] Also included are polypeptide fragments which are differentially modified during or after translation, *e.g.*, by glycosylation, acetylation, phosphorylation, amidation, derivatization by known protecting/blocking groups, proteolytic cleavage, linkage to an antibody molecule or other cellular ligand, and the like. Any of numerous chemical modifications may be carried out by known techniques, including but not limited, to specific chemical cleavage by cyanogen bromide, trypsin, chymotrypsin, papain, V8 protease, NaBH<sub>4</sub>; acetylation, formylation, oxidation, reduction; metabolic synthesis in the presence of tunicamycin; and the like.

[0103] Additional post-translational modifications include, for example, N-linked or O-linked carbohydrate chains, processing of N-terminal or C-terminal ends), attachment of chemical moieties to the amino acid backbone, chemical modifications of N-linked or O-linked carbohydrate chains, and addition or deletion of an N-terminal methionine residue as a result of prokaryotic host cell expression. The polypeptide fragments may also be modified with a detectable label, such as an enzymatic, fluorescent, isotopic or affinity label to allow for detection and isolation of the polypeptide.

[0104] Also provided are chemically modified polypeptide derivatives that may provide additional advantages such as increased solubility, stability and circulating time of the polypeptide, or decreased immunogenicity. See U.S. Pat. No: 4,179,337. The chemical moieties for derivitization may be selected from water soluble polymers such as polyethylene glycol, ethylene glycol/propylene glycol copolymers, carboxymethylcellulose, dextran, polyvinyl alcohol and the like. The polypeptides may be modified at random positions within the molecule, or at predetermined positions within the molecule and may include one, two, three or more attached chemical moieties.

[0105] The polymer may be of any molecular weight, and may be branched or unbranched. For polyethylene glycol, the molecular weight is between about 1 kDa and about 100 kDa (the term "about" indicating that in preparations of polyethylene glycol, some molecules will weigh more, some less, than the stated molecular weight) for ease in handling and manufacturing. Other sizes may be used, depending on the desired therapeutic profile (*e.g.*, the duration of sustained release desired, the effects, if any on biological activity, the ease in handling, the degree or lack of antigenicity and other known effects of the polyethylene glycol to a therapeutic protein or analog).

[0106] The polyethylene glycol molecules (or other chemical moieties) should be attached to the polypeptide with consideration of effects on functional or antigenic domains of the polypeptide. There are a number of attachment methods available to those skilled in the art, *e.g.*, EP 0 401 384, herein incorporated by reference (coupling PEG to G-CSF), see also Malik *et al.* (1992) Exp Hematol. September;20(8):1028-35, reporting pegylation of GM-CSF using tresyl chloride). For example, polyethylene glycol may be covalently bound through amino acid residues via a reactive group, such as, a free amino or carboxyl group. Reactive groups are those to which an activated polyethylene glycol

molecule may be bound. The amino acid residues having a free amino group may include lysine residues and the N-terminal amino acid residues; those having a free carboxyl group may include aspartic acid residues, glutamic acid residues and the C-terminal amino acid residue. Sulfhydryl groups may also be used as a reactive group for attaching the polyethylene glycol molecules. A polymer sometimes is attached at an amino group, such as attachment at the N-terminus or lysine group.

[0107] One may specifically desire proteins chemically modified at the N-terminus. Using polyethylene glycol as an illustration of the present composition, one may select from a variety of polyethylene glycol molecules (by molecular weight, branching, and the like), the proportion of polyethylene glycol molecules to protein (polypeptide) molecules in the reaction mix, the type of pegylation reaction to be performed, and the method of obtaining the selected N-terminally pegylated protein. The method of obtaining the N-terminally pegylated preparation (i.e., separating this moiety from other monopegylated moieties if necessary) may be by purification of the N-terminally pegylated material from a population of pegylated protein molecules. Selective proteins chemically modified at the N-terminus may be accomplished by reductive alkylation, which exploits differential reactivity of different types of primary amino groups (lysine versus the N-terminal) available for derivatization in a particular protein. Under the appropriate reaction conditions, substantially selective derivatization of the protein at the N-terminus with a carbonyl group containing polymer is achieved.

#### Substantially Identical Nucleic Acids and Polypeptides

[0108] Nucleotide sequences and polypeptide sequences that are substantially identical to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence and the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide sequences encoded by those nucleotide sequences are included herein. The term "substantially identical" as used herein refers to two or more nucleic acids or polypeptides sharing one or more identical nucleotide sequences or polypeptide sequences, respectively. Included are nucleotide sequences or polypeptide sequences that are 55% or more, 60% or more, 65% or more, 70% or more, 75% or more, 80% or more, 85% or more, 90% or more, 95% or more (each often within a 1%, 2%, 3% or 4% variability) or more identical to the nucleotide sequences in SEQ ID NO: 1-12 or the encoded *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide amino acid sequences. One test for determining whether two nucleic acids are substantially identical is to determine the percent of identical nucleotide sequences or polypeptide sequences shared between the nucleic acids or polypeptides.

[0109] Calculations of sequence identity are often performed as follows. Sequences are aligned for optimal comparison purposes (e.g., gaps can be introduced in one or both of a first and a second amino acid or nucleic acid sequence for optimal alignment and non-homologous sequences can be disregarded for comparison purposes). The length of a reference sequence aligned for comparison purposes is

sometimes 30% or more, 40% or more, 50% or more, often 60% or more, and more often 70% or more, 80% or more, 90% or more, 90% or more, or 100% of the length of the reference sequence. The nucleotides or amino acids at corresponding nucleotide or polypeptide positions, respectively, are then compared among the two sequences. When a position in the first sequence is occupied by the same nucleotide or amino acid as the corresponding position in the second sequence, the nucleotides or amino acids are deemed to be identical at that position. The percent identity between the two sequences is a function of the number of identical positions shared by the sequences, taking into account the number of gaps, and the length of each gap, introduced for optimal alignment of the two sequences.

[0110] Comparison of sequences and determination of percent identity between two sequences can be accomplished using a mathematical algorithm. Percent identity between two amino acid or nucleotide sequences can be determined using the algorithm of Meyers & Miller, *CABIOS* 4: 11-17 (1989), which has been incorporated into the ALIGN program (version 2.0), using a PAM120 weight residue table, a gap length penalty of 12 and a gap penalty of 4. Also, percent identity between two amino acid sequences can be determined using the Needleman & Wunsch, *J. Mol. Biol.* 48: 444-453 (1970) algorithm which has been incorporated into the GAP program in the GCG software package (available at the http address [www.gcg.com](http://www.gcg.com)), using either a Blossum 62 matrix or a PAM250 matrix, and a gap weight of 16, 14, 12, 10, 8, 6, or 4 and a length weight of 1, 2, 3, 4, 5, or 6. Percent identity between two nucleotide sequences can be determined using the GAP program in the GCG software package (available at http address [www.gcg.com](http://www.gcg.com)), using a NWSgapdna.CMP matrix and a gap weight of 40, 50, 60, 70, or 80 and a length weight of 1, 2, 3, 4, 5, or 6. A set of parameters often used is a Blossum 62 scoring matrix with a gap open penalty of 12, a gap extend penalty of 4, and a frameshift gap penalty of 5.

[0111] Another manner for determining if two nucleic acids are substantially identical is to assess whether a polynucleotide homologous to one nucleic acid will hybridize to the other nucleic acid under stringent conditions. As use herein, the term "stringent conditions" refers to conditions for hybridization and washing. Stringent conditions are known to those skilled in the art and can be found in *Current Protocols in Molecular Biology*, John Wiley & Sons, N.Y. , 6.3.1-6.3.6 (1989). Aqueous and non-aqueous methods are described in that reference and either can be used. An example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 50°C. Another example of stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 55°C. A further example of stringent hybridization conditions is hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one or more washes in 0.2X SSC, 0.1% SDS at 60°C. Often, stringent hybridization conditions are hybridization in 6X sodium chloride/sodium citrate (SSC) at about 45°C, followed by one

or more washes in 0.2X SSC, 0.1% SDS at 65°C. More often, stringency conditions are 0.5M sodium phosphate, 7% SDS at 65°C, followed by one or more washes at 0.2X SSC, 1% SDS at 65°C.

[0112] An example of a substantially identical nucleotide sequence to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence is one that has a different nucleotide sequence but still encodes the same polypeptide sequence encoded by the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence. Another example is a nucleotide sequence that encodes a polypeptide having a polypeptide sequence that is more than 70% or more identical to, sometimes 75% or more, 80% or more, or 85% or more identical to, and often 90% or more and 95% or more identical to a polypeptide sequence encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence.

[0113] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequences and *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* amino acid sequences can be used as “query sequences” to perform a search against public databases to identify other family members or related sequences, for example. Such searches can be performed using the NBLAST and XBLAST programs (version 2.0) of Altschul *et al.*, *J. Mol. Biol.* 215: 403-10 (1990). BLAST nucleotide searches can be performed with the NBLAST program, score = 100, wordlength = 12 to obtain nucleotide sequences homologous to nucleotide sequences from SEQ ID NO: 1-12. BLAST polypeptide searches can be performed with the XBLAST program, score = 50, wordlength = 3 to obtain amino acid sequences homologous to polypeptides encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence. To obtain gapped alignments for comparison purposes, Gapped BLAST can be utilized as described in Altschul *et al.*, *Nucleic Acids Res.* 25(17): 3389-3402 (1997). When utilizing BLAST and Gapped BLAST programs, default parameters of the respective programs (*e.g.*, XBLAST and NBLAST) can be used (*see* the http address [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)).

[0114] A nucleic acid that is substantially identical to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence may include polymorphic sites at positions equivalent to those described herein when the sequences are aligned. For example, using the alignment procedures described herein, SNPs in a sequence substantially identical to a sequence in SEQ ID NO: 1-12 can be identified at nucleotide positions that match (*i.e.*, align) with nucleotides at SNP positions in the nucleotide sequence of SEQ ID NO: 1-12. Also, where a polymorphic variation results in an insertion or deletion, insertion or deletion of a nucleotide sequence from a reference sequence can change the relative positions of other polymorphic sites in the nucleotide sequence.

[0115] Substantially identical nucleotide and polypeptide sequences include those that are naturally occurring, such as allelic variants (same locus), splice variants, homologs (different locus), and orthologs (different organism) or can be non-naturally occurring. Non-naturally occurring variants can be generated by mutagenesis techniques, including those applied to polynucleotides, cells, or organisms.

The variants can contain nucleotide substitutions, deletions, inversions and insertions. Variation can occur in either or both the coding and non-coding regions. The variations can produce both conservative and non-conservative amino acid substitutions (as compared in the encoded product). Orthologs, homologs, allelic variants, and splice variants can be identified using methods known in the art. These variants normally comprise a nucleotide sequence encoding a polypeptide that is 50% or more, about 55% or more, often about 70-75% or more, more often about 80-85% or more, and typically about 90-95% or more identical to the amino acid sequences of target polypeptides or a fragment thereof. Such nucleic acid molecules readily can be identified as being able to hybridize under stringent conditions to a nucleotide sequence in SEQ ID NO: 1-12 or a fragment thereof. Nucleic acid molecules corresponding to orthologs, homologs, and allelic variants of a nucleotide sequence in SEQ ID NO: 1-12 can be identified by mapping the sequence to the same chromosome or locus as the nucleotide sequence in SEQ ID NO: 1-12.

[0116] Also, substantially identical nucleotide sequences may include codons that are altered with respect to the naturally occurring sequence for enhancing expression of a target polypeptide in a particular expression system. For example, the nucleic acid can be one in which one or more codons are altered, and often 10% or more or 20% or more of the codons are altered for optimized expression in bacteria (*e.g.*, *E. coli.*), yeast (*e.g.*, *S. cerevisiae*), human (*e.g.*, 293 cells), insect, or rodent (*e.g.*, hamster) cells.

#### Methods for Identifying Subjects at Risk of Breast Cancer and Breast Cancer Risk in a Subject

[0117] Methods for prognosing and diagnosing breast cancer in subjects are provided herein. These methods include detecting the presence or absence of one or more polymorphic variations associated with breast cancer in a nucleotide sequence set forth in SEQ ID NO: 1-5, or substantially identical sequence thereof, in a sample from a subject, where the presence of a polymorphic variant is indicative of a risk of breast cancer.

[0118] Thus, featured herein is a method for detecting a subject at risk of breast cancer or the risk of breast cancer in a subject, which comprises detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b),



or (c), often a fragment that includes a polymorphic site associated with breast cancer; whereby the presence of the polymorphic variation is indicative of a risk of breast cancer in the subject.

[0119] In certain embodiments, determining the presence of a combination of two or more polymorphic variants associated with breast cancer in one or more genetic loci (e.g., one or more genes) of the sample is determined to identify, quantify and/or estimate, risk of breast cancer. The risk often is the probability of having or developing breast cancer. The risk sometimes is expressed as a relative risk with respect to a population average risk of breast cancer, and sometimes is expressed as a relative risk with respect to the lowest risk group. Such relative risk assessments often are based upon penetrance values determined by statistical methods (see e.g., statistical analysis Example 9), and are particularly useful to clinicians and insurance companies for assessing risk of breast cancer (e.g., a clinician can target appropriate detection, prevention and therapeutic regimens to a patient after determining the patient's risk of breast cancer, and an insurance company can fine tune actuarial tables based upon population genotype assessments of breast cancer risk). Risk of breast cancer sometimes is expressed as an odds ratio, which is the odds of a particular person having a genotype has or will develop breast cancer with respect to another genotype group (e.g., the most disease protective genotype or population average). In related embodiments, the determination is utilized to identify a subject at risk of breast cancer. In an embodiment, two or more polymorphic variations are detected in two or more regions in human genomic DNA associated with increased risk of breast cancer, such as regions selected from the group of loci consisting of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE*, for example. In certain embodiments, 3 or more, or 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 or more polymorphic variants are detected in the sample. In specific embodiments, polymorphic variants are detected in *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and *GALE* loci, such as at positions 44247 in SEQ ID NO: 1 (*ICAM*), position 36424 in SEQ ID NO: 2 (*MAPK10*), position 48563 in SEQ ID NO: 3 (*KIAA0861*), position 49002 in SEQ ID NO: 4 (*NUMA1*) and position 174 in SEQ ID NO: 5 (*GALE*), for example. In certain embodiments, polymorphic variants are detected at other genetic loci (e.g., the polymorphic variants can be detected in *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* and/or *GALE* in addition to other loci or only in other loci), where the other loci include but are not limited to *RAD21*, *KLF12*, *SPUVE*, *GRIN3A*, *PFTK1*, *SERPINA5*, *LOC115209*, *HRMT1L3*, *DLG1*, *KIAA0783*, *DPF3*, *CENPC1*, *GP6*, *LAMA4*, *CHCB/C20ORF154*, *LOC338749*, and *TTN/LOC351327*, which are described in concurrently-filed patent applications having attorney docket numbers 524592006700, 524592006800, 524592007000, 524592007100 and 524592007200, and any others disclosed in patent application nos. 60/429,136 (filed November 25, 2002) 60/490,234 (filed July 24, 2003).

[0120] A risk of developing aggressive forms of breast cancer likely to metastasize or invade surrounding tissues (e.g., Stage IIIA, IIIB, and IV breast cancers), and subjects at risk of developing

aggressive forms of breast cancer also may be identified by the methods described herein. These methods include collecting phenotype information from subjects having breast cancer, which includes the stage of progression of the breast cancer, and performing a secondary phenotype analysis to detect the presence or absence of one or more polymorphic variations associated with a particular stage form of breast cancer. Thus, detecting the presence or absence of one or more polymorphic variations in a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence associated with a late stage form of breast cancer often is prognostic and/or diagnostic of an aggressive form of the cancer.

[0121] Results from prognostic tests may be combined with other test results to diagnose breast cancer. For example, prognostic results may be gathered, a patient sample may be ordered based on a determined predisposition to breast cancer, the patient sample is analyzed, and the results of the analysis may be utilized to diagnose breast cancer. Also breast cancer diagnostic methods can be developed from studies used to generate prognostic/diagnostic methods in which populations are stratified into subpopulations having different progressions of breast cancer. In another embodiment, prognostic results may be gathered; a patient's risk factors for developing breast cancer analyzed (*e.g.*, age, race, family history, age of first menstrual cycle, age at birth of first child); and a patient sample may be ordered based on a determined predisposition to breast cancer. In an alternative embodiment, the results from predisposition analyses described herein may be combined with other test results indicative of breast cancer, which were previously, concurrently, or subsequently gathered with respect to the predisposition testing. In these embodiments, the combination of the prognostic test results with other test results can be probative of breast cancer, and the combination can be utilized as a breast cancer diagnostic. The results of any test indicative of breast cancer known in the art may be combined with the methods described herein. Examples of such tests are mammography (*e.g.*, a more frequent and/or earlier mammography regimen may be prescribed); breast biopsy and optionally a biopsy from another tissue; breast ultrasound and optionally an ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally an MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally of another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and thermal imaging of the breast and optionally of another tissue. Testing may be performed on tissue other than breast to diagnose the occurrence of metastasis (*e.g.*, testing of the lymph node).

[0122] Risk of breast cancer sometimes is expressed as a probability, such as an odds ratio, percentage, or risk factor. The risk is based upon the presence or absence of one or more polymorphic variants described herein, and also may be based in part upon phenotypic traits of the individual being tested. Methods for calculating predispositions based upon patient data are well known (*see, e.g.*, Agresti, *Categorical Data Analysis*, 2nd Ed. 2002. Wiley). Allelotyping and genotyping analyses may be

carried out in populations other than those exemplified herein to enhance the predictive power of the prognostic method. These further analyses are executed in view of the exemplified procedures described herein, and may be based upon the same polymorphic variations or additional polymorphic variations. Risk determinations for breast cancer are useful in a variety of applications. In one embodiment, breast cancer risk determinations are used by clinicians to direct appropriate detection, preventative and treatment procedures to subjects who most require these. In another embodiment, breast cancer risk determinations are used by health insurers for preparing actuarial tables and for calculating insurance premiums.

[0123] The nucleic acid sample typically is isolated from a biological sample obtained from a subject. For example, nucleic acid can be isolated from blood, saliva, sputum, urine, cell scrapings, and biopsy tissue. The nucleic acid sample can be isolated from a biological sample using standard techniques, such as the technique described in Example 2. As used herein, the term “subject” refers primarily to humans but also refers to other mammals such as dogs, cats, and ungulates (*e.g.*, cattle, sheep, and swine). Subjects also include avians (*e.g.*, chickens and turkeys), reptiles, and fish (*e.g.*, salmon), as embodiments described herein can be adapted to nucleic acid samples isolated from any of these organisms. The nucleic acid sample may be isolated from the subject and then directly utilized in a method for determining the presence of a polymorphic variant, or alternatively, the sample may be isolated and then stored (*e.g.*, frozen) for a period of time before being subjected to analysis.

[0124] The presence or absence of a polymorphic variant is determined using one or both chromosomal complements represented in the nucleic acid sample. Determining the presence or absence of a polymorphic variant in both chromosomal complements represented in a nucleic acid sample from a subject having a copy of each chromosome is useful for determining the zygosity of an individual for the polymorphic variant (*i.e.*, whether the individual is homozygous or heterozygous for the polymorphic variant). Any oligonucleotide-based diagnostic may be utilized to determine whether a sample includes the presence or absence of a polymorphic variant in a sample. For example, primer extension methods, ligase sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,679,524 and 5,952,174, and WO 01/27326), mismatch sequence determination methods (*e.g.*, U.S. Pat. Nos. 5,851,770; 5,958,692; 6,110,684; and 6,183,958), microarray sequence determination methods, restriction fragment length polymorphism (RFLP), single strand conformation polymorphism detection (SSCP) (*e.g.*, U.S. Pat. Nos. 5,891,625 and 6,013,499), PCR-based assays (*e.g.*, TAQMAN<sup>®</sup> PCR System (Applied Biosystems)), and nucleotide sequencing methods may be used.

[0125] Oligonucleotide extension methods typically involve providing a pair of oligonucleotide primers in a polymerase chain reaction (PCR) or in other nucleic acid amplification methods for the purpose of amplifying a region from the nucleic acid sample that comprises the polymorphic variation.

One oligonucleotide primer is complementary to a region 3' of the polymorphism and the other is complementary to a region 5' of the polymorphism. A PCR primer pair may be used in methods disclosed in U.S. Pat. Nos. 4,683,195; 4,683,202, 4,965,188; 5,656,493; 5,998,143; 6,140,054; WO 01/27327; and WO 01/27329 for example. PCR primer pairs may also be used in any commercially available machines that perform PCR, such as any of the GENEAMP® Systems available from Applied Biosystems. Also, those of ordinary skill in the art will be able to design oligonucleotide primers based upon a nucleotide sequence set forth in SEQ ID NO: 1-5 without undue experimentation using knowledge readily available in the art.

[0126] Also provided is an extension oligonucleotide that hybridizes to the amplified fragment adjacent to the polymorphic variation. As used herein, the term "adjacent" refers to the 3' end of the extension oligonucleotide being often 1 nucleotide from the 5' end of the polymorphic site, and sometimes 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides from the 5' end of the polymorphic site, in the nucleic acid when the extension oligonucleotide is hybridized to the nucleic acid. The extension oligonucleotide then is extended by one or more nucleotides, and the number and/or type of nucleotides that are added to the extension oligonucleotide determine whether the polymorphic variant is present. Oligonucleotide extension methods are disclosed, for example, in U.S. Pat. Nos. 4,656,127; 4,851,331; 5,679,524; 5,834,189; 5,876,934; 5,908,755; 5,912,118; 5,976,802; 5,981,186; 6,004,744; 6,013,431; 6,017,702; 6,046,005; 6,087,095; 6,210,891; and WO 01/20039. Oligonucleotide extension methods using mass spectrometry are described, for example, in U.S. Pat. Nos. 5,547,835; 5,605,798; 5,691,141; 5,849,542; 5,869,242; 5,928,906; 6,043,031; and 6,194,144, and a method often utilized is described herein in Example 2. Multiple extension oligonucleotides may be utilized in one reaction, which is referred to herein as "multiplexing."

[0127] A microarray can be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A microarray may include any oligonucleotides described herein, and methods for making and using oligonucleotide microarrays suitable for diagnostic use are disclosed in U.S. Pat. Nos. 5,492,806; 5,525,464; 5,589,330; 5,695,940; 5,849,483; 6,018,041; 6,045,996; 6,136,541; 6,142,681; 6,156,501; 6,197,506; 6,223,127; 6,225,625; 6,229,911; 6,239,273; WO 00/52625; WO 01/25485; and WO 01/29259. The microarray typically comprises a solid support and the oligonucleotides may be linked to this solid support by covalent bonds or by non-covalent interactions. The oligonucleotides may also be linked to the solid support directly or by a spacer molecule. A microarray may comprise one or more oligonucleotides complementary to a polymorphic site set forth in SEQ ID NO: 1-5 or below.

[0128] A kit also may be utilized for determining whether a polymorphic variant is present or absent in a nucleic acid sample. A kit often comprises one or more pairs of oligonucleotide primers useful for

amplifying a fragment of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence or a substantially identical sequence thereof, where the fragment includes a polymorphic site. The kit sometimes comprises a polymerizing agent, for example, a thermostable nucleic acid polymerase such as one disclosed in U.S. Pat. Nos. 4,889,818 or 6,077,664. Also, the kit often comprises an elongation oligonucleotide that hybridizes to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence in a nucleic acid sample adjacent to the polymorphic site. Where the kit includes an elongation oligonucleotide, it also often comprises chain elongating nucleotides, such as dATP, dTTP, dGTP, dCTP, and dITP, including analogs of dATP, dTTP, dGTP, dCTP and dITP, provided that such analogs are substrates for a thermostable nucleic acid polymerase and can be incorporated into a nucleic acid chain elongated from the extension oligonucleotide. Along with chain elongating nucleotides would be one or more chain terminating nucleotides such as ddATP, ddTTP, ddGTP, ddCTP, and the like. In an embodiment, the kit comprises one or more oligonucleotide primer pairs, a polymerizing agent, chain elongating nucleotides, at least one elongation oligonucleotide, and one or more chain terminating nucleotides. Kits optionally include buffers, vials, microtiter plates, and instructions for use.

[0129] An individual identified as being at risk of breast cancer may be heterozygous or homozygous with respect to the allele associated with a higher risk of breast cancer. A subject homozygous for an allele associated with an increased risk of breast cancer is at a comparatively high risk of breast cancer, a subject heterozygous for an allele associated with an increased risk of breast cancer is at a comparatively intermediate risk of breast cancer, and a subject homozygous for an allele associated with a decreased risk of breast cancer is at a comparatively low risk of breast cancer. A genotype may be assessed for a complementary strand, such that the complementary nucleotide at a particular position is detected.

[0130] Also featured are methods for determining risk of breast cancer and/or identifying a subject at risk of breast cancer by contacting a polypeptide or protein encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence from a subject with an antibody that specifically binds to an epitope associated with increased risk of breast cancer in the polypeptide. In certain embodiments, the antibody specifically binds to an epitope that comprises a leucine at amino acid position 359 in SEQ ID NO: 17, a leucine at amino acid position 378 in SEQ ID NO: 17, or an alanine at amino acid position 857 in SEQ ID NO: 17, a proline at amino acid position 352 in SEQ ID NO: 15 or an alanine at amino acid position 348 in SEQ ID NO: 15.

#### Applications of Prognostic and Diagnostic Results to Pharmacogenomic Methods

[0131] Pharmacogenomics is a discipline that involves tailoring a treatment for a subject according to the subject's genotype. For example, based upon the outcome of a prognostic test described herein, a

clinician or physician may target pertinent information and preventative or therapeutic treatments to a subject who would be benefited by the information or treatment and avoid directing such information and treatments to a subject who would not be benefited (*e.g.*, the treatment has no therapeutic effect and/or the subject experiences adverse side effects). As therapeutic approaches for breast cancer continue to evolve and improve, the goal of treatments for breast cancer related disorders is to intervene even before clinical signs (*e.g.*, identification of lump in the breast) first manifest. Thus, genetic markers associated with susceptibility to breast cancer prove useful for early diagnosis, prevention and treatment of breast cancer.

[0132] The following is an example of a pharmacogenomic embodiment. A particular treatment regimen can exert a differential effect depending upon the subject's genotype. Where a candidate therapeutic exhibits a significant interaction with a major allele and a comparatively weak interaction with a minor allele (*e.g.*, an order of magnitude or greater difference in the interaction), such a therapeutic typically would not be administered to a subject genotyped as being homozygous for the minor allele, and sometimes not administered to a subject genotyped as being heterozygous for the minor allele. In another example, where a candidate therapeutic is not significantly toxic when administered to subjects who are homozygous for a major allele but is comparatively toxic when administered to subjects heterozygous or homozygous for a minor allele, the candidate therapeutic is not typically administered to subjects who are genotyped as being heterozygous or homozygous with respect to the minor allele.

[0133] The methods described herein are applicable to pharmacogenomic methods for detecting, preventing, alleviating and/or treating breast cancer. For example, a nucleic acid sample from an individual may be subjected to a genetic test described herein. Where one or more polymorphic variations associated with increased risk of breast cancer are identified in a subject, information for detecting, preventing or treating breast cancer and/or one or more breast cancer detection, prevention and/or treatment regimens then may be directed to and/or prescribed to that subject.

[0134] In certain embodiments, a detection, preventive and/or treatment regimen is specifically prescribed and/or administered to individuals who will most benefit from it based upon their risk of developing breast cancer assessed by the methods described herein. Thus, provided are methods for identifying a subject at risk of breast cancer and then prescribing a detection, therapeutic or preventative regimen to individuals identified as being at risk of breast cancer. Thus, certain embodiments are directed to methods for treating breast cancer in a subject, reducing risk of breast cancer in a subject, or early detection of breast cancer in a subject, which comprise: detecting the presence or absence of a polymorphic variant associated with breast cancer in a nucleotide sequence in a nucleic acid sample from a subject, where the nucleotide sequence comprises a polynucleotide sequence selected from the group consisting of: (a) a nucleotide sequence set forth in SEQ ID NO: 1-5; (b) a nucleotide sequence which

encodes a polypeptide having an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5; (c) a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5 or a nucleotide sequence about 90% or more identical to the nucleotide sequence set forth in SEQ ID NO: 1-5; and (d) a fragment of a nucleotide sequence of (a), (b), or (c), sometimes comprising a polymorphic site associated with breast cancer; and prescribing or administering a breast cancer treatment regimen, preventative regimen and/or detection regimen to a subject from whom the sample originated where the presence of one or more polymorphic variations associated with breast cancer are detected in the nucleotide sequence. In these methods, genetic results may be utilized in combination with other test results to diagnose breast cancer as described above. Other test results include but are not limited to mammography results, imaging results, biopsy results and results from *BRCA1* or *BRAC2* test results, as described above.

[0135] Detection regimens include one or more mammography procedures, a regular mammography regimen (*e.g.*, once a year, or once every six, four, three or two months); an early mammography regimen (*e.g.*, mammography tests are performed beginning at age 25, 30, or 35); one or more biopsy procedures (*e.g.*, a regular biopsy regimen beginning at age 40); breast biopsy and biopsy from other tissue; breast ultrasound and optionally ultrasound analysis of another tissue; breast magnetic resonance imaging (MRI) and optionally MRI analysis of another tissue; electrical impedance (T-scan) analysis of breast and optionally another tissue; ductal lavage; nuclear medicine analysis (*e.g.*, scintimammography); *BRCA1* and/or *BRCA2* sequence analysis results; and/or thermal imaging of the breast and optionally another tissue.

[0136] Treatments sometimes are preventative (*e.g.*, is prescribed or administered to reduce the probability that a breast cancer associated condition arises or progresses), sometimes are therapeutic, and sometimes delay, alleviate or halt the progression of breast cancer. Any known preventative or therapeutic treatment for alleviating or preventing the occurrence of breast cancer is prescribed and/or administered. For example, certain preventative treatments often are prescribed to subjects having a predisposition to breast cancer and where the subject is not diagnosed with breast cancer or is diagnosed as having symptoms indicative of early stage breast cancer (*e.g.*, stage I). For subjects not diagnosed as having breast cancer, any preventative treatments known in the art can be prescribed and administered, which include selective hormone receptor modulators (*e.g.*, selective estrogen receptor modulators (SERMs) such as tamoxifen, reloxifene, and toremifene); compositions that prevent production of hormones (*e.g.*, aromatase inhibitors that prevent the production of estrogen in the adrenal gland, such as exemestane, letrozole, anastrozol, goserelin, and megestrol); other hormonal treatments (*e.g.*, goserelin acetate and fulvestrant); biologic response modifiers such as antibodies (*e.g.*, trastuzumab (herceptin/HER2)); surgery (*e.g.*, lumpectomy and mastectomy); drugs that delay or halt metastasis (*e.g.*,

pamidronate disodium); and alternative/complementary medicine (*e.g.*, acupuncture, acupressure, moxibustion, qi gong, reiki, ayurveda, vitamins, minerals, and herbs (*e.g.*, astragalus root, burdock root, garlic, green tea, and licorice root)).

[0137] The use of breast cancer treatments are well known in the art, and include surgery, chemotherapy and/or radiation therapy. Any of the treatments may be used in combination to treat or prevent breast cancer (*e.g.*, surgery followed by radiation therapy or chemotherapy). Examples of chemotherapy combinations used to treat breast cancer include: cyclophosphamide (Cytoxan), methotrexate (Amethopterin, Mexate, Folex), and fluorouracil (Fluorouracil, 5-Fu, Adrucil), which is referred to as CMF; cyclophosphamide, doxorubicin (Adriamycin), and fluorouracil, which is referred to as CAF; and doxorubicin (Adriamycin) and cyclophosphamide, which is referred to as AC.

[0138] As breast cancer preventative and treatment information can be specifically targeted to subjects in need thereof (*e.g.*, those at risk of developing breast cancer or those that have early signs of breast cancer), provided herein is a method for preventing or reducing the risk of developing breast cancer in a subject, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying a subject with a predisposition to breast cancer, whereby the presence of the polymorphic variation is indicative of a predisposition to breast cancer in the subject; and (c) if such a predisposition is identified, providing the subject with information about methods or products to prevent or reduce breast cancer or to delay the onset of breast cancer. Also provided is a method of targeting information or advertising to a subpopulation of a human population based on the subpopulation being genetically predisposed to a disease or condition, which comprises: (a) detecting the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) identifying the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; and (c) providing information only to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition.

[0139] Pharmacogenomics methods also may be used to analyze and predict a response to a breast cancer treatment or a drug. For example, if pharmacogenomics analysis indicates a likelihood that an individual will respond positively to a breast cancer treatment with a particular drug, the drug may be administered to the individual. Conversely, if the analysis indicates that an individual is likely to respond negatively to treatment with a particular drug, an alternative course of treatment may be prescribed. A negative response may be defined as either the absence of an efficacious response or the presence of toxic side effects. The response to a therapeutic treatment can be predicted in a background study in which subjects in any of the following populations are genotyped: a population that responds favorably to a



treatment regimen, a population that does not respond significantly to a treatment regimen, and a population that responds adversely to a treatment regimen (*e.g.*, exhibits one or more side effects). These populations are provided as examples and other populations and subpopulations may be analyzed. Based upon the results of these analyses, a subject is genotyped to predict whether he or she will respond favorably to a treatment regimen, not respond significantly to a treatment regimen, or respond adversely to a treatment regimen.

[0140] The methods described herein also are applicable to clinical drug trials. One or more polymorphic variants indicative of response to an agent for treating breast cancer or to side effects to an agent for treating breast cancer may be identified using the methods described herein. Thereafter, potential participants in clinical trials of such an agent may be screened to identify those individuals most likely to respond favorably to the drug and exclude those likely to experience side effects. In that way, the effectiveness of drug treatment may be measured in individuals who respond positively to the drug, without lowering the measurement as a result of the inclusion of individuals who are unlikely to respond positively in the study and without risking undesirable safety problems. In certain embodiments, the agent for treating breast cancer described herein targets *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* or a target in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* pathway.

[0141] Thus, another embodiment is a method of selecting an individual for inclusion in a clinical trial of a treatment or drug comprising the steps of: (a) obtaining a nucleic acid sample from an individual; (b) determining the identity of a polymorphic variation which is associated with a positive response to the treatment or the drug, or at least one polymorphic variation which is associated with a negative response to the treatment or the drug in the nucleic acid sample, and (c) including the individual in the clinical trial if the nucleic acid sample contains said polymorphic variation associated with a positive response to the treatment or the drug or if the nucleic acid sample lacks said polymorphic variation associated with a negative response to the treatment or the drug. In addition, the methods for selecting an individual for inclusion in a clinical trial of a treatment or drug encompass methods with any further limitation described in this disclosure, or those following, specified alone or in any combination. The polymorphic variation may be in a sequence selected individually or in any combination from the group consisting of (i) a polynucleotide sequence set forth in SEQ ID NO: 1-5; (ii) a polynucleotide sequence that is 90% or more identical to a nucleotide sequence set forth in SEQ ID NO: 1-5; (iii) a polynucleotide sequence that encodes a polypeptide having an amino acid sequence identical to or 90% or more identical to an amino acid sequence encoded by a nucleotide sequence set forth in SEQ ID NO: 1-5; and (iv) a fragment of a polynucleotide sequence of (i), (ii), or (iii) comprising the polymorphic site. The including step (c) optionally comprises administering the drug or the treatment to the individual if the nucleic acid sample contains the polymorphic variation associated with a positive response to the

treatment or the drug and the nucleic acid sample lacks said biallelic marker associated with a negative response to the treatment or the drug.

[0142] Also provided herein is a method of partnering between a diagnostic/prognostic testing provider and a provider of a consumable product, which comprises: (a) the diagnostic/prognostic testing provider detects the presence or absence of a polymorphic variation associated with breast cancer at a polymorphic site in a nucleotide sequence in a nucleic acid sample from a subject; (b) the diagnostic/prognostic testing provider identifies the subpopulation of subjects in which the polymorphic variation is associated with breast cancer; (c) the diagnostic/prognostic testing provider forwards information to the subpopulation of subjects about a particular product which may be obtained and consumed or applied by the subject to help prevent or delay onset of the disease or condition; and (d) the provider of a consumable product forwards to the diagnostic test provider a fee every time the diagnostic/prognostic test provider forwards information to the subject as set forth in step (c) above.

#### Compositions Comprising Breast Cancer-Directed Molecules

[0143] Featured herein is a composition comprising a breast cancer cell and one or more molecules specifically directed and targeted to a nucleic acid comprising a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence or a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Such directed molecules include, but are not limited to, a compound that binds to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid or a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide; a RNAi or siRNA molecule having a strand complementary to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence; an antisense nucleic acid complementary to an RNA encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* DNA sequence; a ribozyme that hybridizes to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence; a nucleic acid aptamer that specifically binds a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide; and an antibody that specifically binds to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or binds to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid. In certain embodiments, the antibody specifically binds to an epitope that comprises a leucine at amino acid position 359 in SEQ ID NO: 17, a leucine at amino acid position 378 in SEQ ID NO: 17, or an alanine at amino acid position 857 in SEQ ID NO: 17, a proline at amino acid position 352 in SEQ ID NO: 15 or an alanine at amino acid position 348 in SEQ ID NO: 15. In specific embodiments, the breast cancer directed molecule interacts with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid or polypeptide variant associated with breast cancer. In other embodiments, the breast cancer directed molecule interacts with a polypeptide involved in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* signal pathway, or a nucleic acid encoding such a

polypeptide. Polypeptides involved in the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* signal pathway are discussed herein.

[0144] Compositions sometimes include an adjuvant known to stimulate an immune response, and in certain embodiments, an adjuvant that stimulates a T-cell lymphocyte response. Adjuvants are known, including but not limited to an aluminum adjuvant (*e.g.*, aluminum hydroxide); a cytokine adjuvant or adjuvant that stimulates a cytokine response (*e.g.*, interleukin (IL)-12 and/or  $\gamma$ -interferon cytokines); a Freund-type mineral oil adjuvant emulsion (*e.g.*, Freund's complete or incomplete adjuvant); a synthetic lipoid compound; a copolymer adjuvant (*e.g.*, TitreMax); a saponin; Quil A; a liposome; an oil-in-water emulsion (*e.g.*, an emulsion stabilized by Tween 80 and pluronic polyoxyethylene/polyoxypropylene block copolymer (Syntex Adjuvant Formulation); TitreMax; detoxified endotoxin (MPL) and mycobacterial cell wall components (TDW, CWS) in 2% squalene (Ribi Adjuvant System)); a muramyl dipeptide; an immune-stimulating complex (ISCOM, *e.g.*, an Ag-modified saponin/cholesterol micelle that forms stable cage-like structure); an aqueous phase adjuvant that does not have a depot effect (*e.g.*, Gerbu adjuvant); a carbohydrate polymer (*e.g.*, AdjuPrime); L-tyrosine; a manide-oleate compound (*e.g.*, Montanide); an ethylene-vinyl acetate copolymer (*e.g.*, Elvax 40W1,2); or lipid A, for example. Such compositions are useful for generating an immune response against a breast cancer directed molecule (*e.g.*, an HLA-binding subsequence within a polypeptide encoded by a nucleotide sequence in SEQ ID NO: 1). In such methods, a peptide having an amino acid subsequence of a polypeptide encoded by a nucleotide sequence in SEQ ID NO: 1-5 is delivered to a subject, where the subsequence binds to an HLA molecule and induces a CTL lymphocyte response. The peptide sometimes is delivered to the subject as an isolated peptide or as a minigene in a plasmid that encodes the peptide. Methods for identifying HLA-binding subsequences in such polypeptides are known (*see e.g.*, publication WO02/20616 and PCT application US98/01373 for methods of identifying such sequences).

[0145] The breast cancer cell may be in a group of breast cancer cells and/or other types of cells cultured *in vitro* or in a tissue having breast cancer cells (*e.g.*, a melanocytic lesion) maintained *in vitro* or present in an animal *in vivo* (*e.g.*, a rat, mouse, ape or human). In certain embodiments, a composition comprises a component from a breast cancer cell or from a subject having a breast cancer cell instead of the breast cancer cell or in addition to the breast cancer cell, where the component sometimes is a nucleic acid molecule (*e.g.*, genomic DNA), a protein mixture or isolated protein, for example. The aforementioned compositions have utility in diagnostic, prognostic and pharmacogenomic methods described previously and in breast cancer therapeutics described hereafter. Certain breast cancer molecules are described in greater detail below.

Compounds

[0146] Compounds can be obtained using any of the numerous approaches in combinatorial library methods known in the art, including: biological libraries; peptoid libraries (libraries of molecules having the functionalities of peptides, but with a novel, non-peptide backbone which are resistant to enzymatic degradation but which nevertheless remain bioactive (see, *e.g.*, Zuckermann *et al.*, J. Med. Chem. 37: 2678-85 (1994)); spatially addressable parallel solid phase or solution phase libraries; synthetic library methods requiring deconvolution; "one-bead one-compound" library methods; and synthetic library methods using affinity chromatography selection. Biological library and peptoid library approaches are typically limited to peptide libraries, while the other approaches are applicable to peptide, non-peptide oligomer or small molecule libraries of compounds (Lam, Anticancer Drug Des. 12: 145, (1997)). Examples of methods for synthesizing molecular libraries are described, for example, in DeWitt *et al.*, Proc. Natl. Acad. Sci. U.S.A. 90: 6909 (1993); Erb *et al.*, Proc. Natl. Acad. Sci. USA 91: 11422 (1994); Zuckermann *et al.*, J. Med. Chem. 37: 2678 (1994); Cho *et al.*, Science 261: 1303 (1993); Carrell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2059 (1994); Carell *et al.*, Angew. Chem. Int. Ed. Engl. 33: 2061 (1994); and in Gallop *et al.*, J. Med. Chem. 37: 1233 (1994).

[0147] Libraries of compounds may be presented in solution (*e.g.*, Houghten, Biotechniques 13: 412-421 (1992)), or on beads (Lam, Nature 354: 82-84 (1991)), chips (Fodor, Nature 364: 555-556 (1993)), bacteria or spores (Ladner, United States Patent No. 5,223,409), plasmids (Cull *et al.*, Proc. Natl. Acad. Sci. USA 89: 1865-1869 (1992)) or on phage (Scott and Smith, Science 249: 386-390 (1990); Devlin, Science 249: 404-406 (1990); Cwirla *et al.*, Proc. Natl. Acad. Sci. 87: 6378-6382 (1990); Felici, J. Mol. Biol. 222: 301-310 (1991); Ladner *supra.*).

[0148] A compound sometimes alters expression and sometimes alters activity of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide and may be a small molecule. Small molecules include, but are not limited to, peptides, peptidomimetics (*e.g.*, peptoids), amino acids, amino acid analogs, polynucleotides, polynucleotide analogs, nucleotides, nucleotide analogs, organic or inorganic compounds (*i.e.*, including heteroorganic and organometallic compounds) having a molecular weight less than about 10,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 5,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 1,000 grams per mole, organic or inorganic compounds having a molecular weight less than about 500 grams per mole, and salts, esters, and other pharmaceutically acceptable forms of such compounds.

Antisense Nucleic Acid Molecules, Ribozymes, RNAi, siRNA and Modified Nucleic Acid Molecules

[0149] An “antisense” nucleic acid refers to a nucleotide sequence complementary to a “sense” nucleic acid encoding a polypeptide, *e.g.*, complementary to the coding strand of a double-stranded cDNA molecule or complementary to an mRNA sequence. The antisense nucleic acid can be complementary to an entire coding strand in SEQ ID NO: 1-12, or to a portion thereof or a substantially identical sequence thereof. In another embodiment, the antisense nucleic acid molecule is antisense to a “noncoding region” of the coding strand of a nucleotide sequence in SEQ ID NO: 1-12 (*e.g.*, 5’ and 3’ untranslated regions).

[0150] An antisense nucleic acid can be designed such that it is complementary to the entire coding region of an mRNA encoded by a nucleotide sequence in SEQ ID NO: 1-4 (*e.g.*, SEQ ID NO: 6-12), and often the antisense nucleic acid is an oligonucleotide antisense to only a portion of a coding or noncoding region of the mRNA. For example, the antisense oligonucleotide can be complementary to the region surrounding the translation start site of the mRNA, *e.g.*, between the -10 and +10 regions of the target gene nucleotide sequence of interest. An antisense oligonucleotide can be, for example, about 7, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, or more nucleotides in length. The antisense nucleic acids, which include the ribozymes described hereafter, can be designed to target a nucleotide sequence in SEQ ID NO: 1-12, often a variant associated with breast cancer, or a substantially identical sequence thereof. Among the variants, minor alleles and major alleles can be targeted, and those associated with a higher risk of breast cancer are often designed, tested, and administered to subjects.

[0151] An antisense nucleic acid can be constructed using chemical synthesis and enzymatic ligation reactions using standard procedures. For example, an antisense nucleic acid (*e.g.*, an antisense oligonucleotide) can be chemically synthesized using naturally occurring nucleotides or variously modified nucleotides designed to increase the biological stability of the molecules or to increase the physical stability of the duplex formed between the antisense and sense nucleic acids, *e.g.*, phosphorothioate derivatives and acridine substituted nucleotides can be used. Antisense nucleic acid also can be produced biologically using an expression vector into which a nucleic acid has been subcloned in an antisense orientation (*i.e.*, RNA transcribed from the inserted nucleic acid will be of an antisense orientation to a target nucleic acid of interest, described further in the following subsection).

[0152] When utilized as therapeutics, antisense nucleic acids typically are administered to a subject (*e.g.*, by direct injection at a tissue site) or generated *in situ* such that they hybridize with or bind to cellular mRNA and/or genomic DNA encoding a polypeptide and thereby inhibit expression of the polypeptide, for example, by inhibiting transcription and/or translation. Alternatively, antisense nucleic acid molecules can be modified to target selected cells and then are administered systemically. For

systemic administration, antisense molecules can be modified such that they specifically bind to receptors or antigens expressed on a selected cell surface, for example, by linking antisense nucleic acid molecules to peptides or antibodies which bind to cell surface receptors or antigens. Antisense nucleic acid molecules can also be delivered to cells using the vectors described herein. Sufficient intracellular concentrations of antisense molecules are achieved by incorporating a strong promoter, such as a pol II or pol III promoter, in the vector construct.

[0153] Antisense nucleic acid molecules sometimes are \*-anomeric nucleic acid molecules. An \*-anomeric nucleic acid molecule forms specific double-stranded hybrids with complementary RNA in which, contrary to the usual \*-units, the strands run parallel to each other (Gaultier *et al.*, Nucleic Acids. Res. 15: 6625-6641 (1987)). Antisense nucleic acid molecules can also comprise a 2'-o-methylribonucleotide (Inoue *et al.*, Nucleic Acids Res. 15: 6131-6148 (1987)) or a chimeric RNA-DNA analogue (Inoue *et al.*, FEBS Lett. 215: 327-330 (1987)). Antisense nucleic acids sometimes are composed of DNA or PNA or any other nucleic acid derivatives described previously.

[0154] In another embodiment, an antisense nucleic acid is a ribozyme. A ribozyme having specificity for a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence can include one or more sequences complementary to such a nucleotide sequence, and a sequence having a known catalytic region responsible for mRNA cleavage (see *e.g.*, U.S. Pat. No. 5,093,246 or Haselhoff and Gerlach, Nature 334: 585-591 (1988)). For example, a derivative of a Tetrahymena L-19 IVS RNA is sometimes utilized in which the nucleotide sequence of the active site is complementary to the nucleotide sequence to be cleaved in a mRNA (see *e.g.*, Cech *et al.* U.S. Patent No. 4,987,071; and Cech *et al.* U.S. Patent No. 5,116,742). Also, target mRNA sequences can be used to select a catalytic RNA having a specific ribonuclease activity from a pool of RNA molecules (see *e.g.*, Bartel & Szostak, Science 261: 1411-1418 (1993)).

[0155] Breast cancer directed molecules include in certain embodiments nucleic acids that can form triple helix structures with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence or a substantially identical sequence thereof, especially one that includes a regulatory region that controls expression of a polypeptide. Gene expression can be inhibited by targeting nucleotide sequences complementary to the regulatory region of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence or a substantially identical sequence (*e.g.*, promoter and/or enhancers) to form triple helical structures that prevent transcription of a gene in target cells (see *e.g.*, Helene, Anticancer Drug Des. 6(6): 569-84 (1991); Helene *et al.*, Ann. N.Y. Acad. Sci. 660: 27-36 (1992); and Maher, Bioassays 14(12): 807-15 (1992). Potential sequences that can be targeted for triple helix formation can be increased by creating a so-called "switchback" nucleic acid molecule. Switchback molecules are synthesized in an alternating 5'-3', 3'-5' manner, such that they base pair with first one strand of a duplex and then the

other, eliminating the necessity for a sizeable stretch of either purines or pyrimidines to be present on one strand of a duplex.

[0156] Breast cancer directed molecules include RNAi and siRNA nucleic acids. Gene expression may be inhibited by the introduction of double-stranded RNA (dsRNA), which induces potent and specific gene silencing, a phenomenon called RNA interference or RNAi. See, *e.g.*, Fire *et al.*, US Patent Number 6,506,559; Tuschl *et al.* PCT International Publication No. WO 01/75164; Kay *et al.* PCT International Publication No. WO 03/010180A1; or Bosher JM, Labouesse, Nat Cell Biol 2000 Feb;2(2):E31-6. This process has been improved by decreasing the size of the double-stranded RNA to 20-24 base pairs (to create small-interfering RNAs or siRNAs) that “switched off” genes in mammalian cells without initiating an acute phase response, *i.e.*, a host defense mechanism that often results in cell death (see, *e.g.*, Caplen *et al.* Proc Natl Acad Sci U S A. 2001 Aug 14;98(17):9742-7 and Elbashir *et al.* Methods 2002 Feb;26(2):199-213). There is increasing evidence of post-transcriptional gene silencing by RNA interference (RNAi) for inhibiting targeted expression in mammalian cells at the mRNA level, in human cells. There is additional evidence of effective methods for inhibiting the proliferation and migration of tumor cells in human patients, and for inhibiting metastatic cancer development (see, *e.g.*, U.S. Patent Application No. US2001000993183; Caplen *et al.* Proc Natl Acad Sci U S A; and Abderrahmani *et al.* Mol Cell Biol 2001 Nov21(21):7256-67).

[0157] An “siRNA” or “RNAi” refers to a nucleic acid that forms a double stranded RNA and has the ability to reduce or inhibit expression of a gene or target gene when the siRNA is delivered to or expressed in the same cell as the gene or target gene. “siRNA” refers to short double-stranded RNA formed by the complementary strands. Complementary portions of the siRNA that hybridize to form the double stranded molecule often have substantial or complete identity to the target molecule sequence. In one embodiment, an siRNA refers to a nucleic acid that has substantial or complete identity to a target gene and forms a double stranded siRNA.

[0158] When designing the siRNA molecules, the targeted region often is selected from a given DNA sequence beginning 50 to 100 nucleotides downstream of the start codon. See, *e.g.*, Elbashir *et al.*, Methods 26:199-213 (2002). Initially, 5’ or 3’ UTRs and regions nearby the start codon were avoided assuming that UTR-binding proteins and/or translation initiation complexes may interfere with binding of the siRNP or RISC endonuclease complex. Sometimes regions of the target 23 nucleotides in length conforming to the sequence motif AA(N19)TT (N, an nucleotide), and regions with approximately 30% to 70% G/C-content (often about 50% G/C-content) often are selected. If no suitable sequences are found, the search often is extended using the motif NA(N21). The sequence of the sense siRNA sometimes corresponds to (N19) TT or N21 (position 3 to 23 of the 23-nt motif), respectively. In the latter case, the 3’ end of the sense siRNA often is converted to TT. The rationale for this sequence

conversion is to generate a symmetric duplex with respect to the sequence composition of the sense and antisense 3' overhangs. The antisense siRNA is synthesized as the complement to position 1 to 21 of the 23-nt motif. Because position 1 of the 23-nt motif is not recognized sequence-specifically by the antisense siRNA, the 3'-most nucleotide residue of the antisense siRNA can be chosen deliberately. However, the penultimate nucleotide of the antisense siRNA (complementary to position 2 of the 23-nt motif) often is complementary to the targeted sequence. For simplifying chemical synthesis, TT often is utilized. siRNAs corresponding to the target motif NAR(N17)YNN, where R is purine (A,G) and Y is pyrimidine (C,U), often are selected. Respective 21 nucleotide sense and antisense siRNAs often begin with a purine nucleotide and can also be expressed from pol III expression vectors without a change in targeting site. Expression of RNAs from pol III promoters often is efficient when the first transcribed nucleotide is a purine.

**[0159]** The sequence of the siRNA can correspond to the full length target gene, or a subsequence thereof. Often, the siRNA is about 15 to about 50 nucleotides in length (*e.g.*, each complementary sequence of the double stranded siRNA is 15-50 nucleotides in length, and the double stranded siRNA is about 15-50 base pairs in length, sometimes about 20-30 nucleotides in length or about 20-25 nucleotides in length, *e.g.*, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, or 30 nucleotides in length. The siRNA sometimes is about 21 nucleotides in length. Methods of using siRNA are well known in the art, and specific siRNA molecules may be purchased from a number of companies including Dharmacon Research, Inc.

**[0160]** Antisense, ribozyme, RNAi and siRNA nucleic acids can be altered to form modified nucleic acid molecules. The nucleic acids can be altered at base moieties, sugar moieties or phosphate backbone moieties to improve stability, hybridization, or solubility of the molecule. For example, the deoxyribose phosphate backbone of nucleic acid molecules can be modified to generate peptide nucleic acids (see Hyrup *et al.*, *Bioorganic & Medicinal Chemistry* 4 (1): 5-23 (1996)). As used herein, the terms "peptide nucleic acid" or "PNA" refers to a nucleic acid mimic such as a DNA mimic, in which the deoxyribose phosphate backbone is replaced by a pseudopeptide backbone and only the four natural nucleobases are retained. The neutral backbone of a PNA can allow for specific hybridization to DNA and RNA under conditions of low ionic strength. Synthesis of PNA oligomers can be performed using standard solid phase peptide synthesis protocols as described, for example, in Hyrup *et al.*, (1996) *supra* and Perry-O'Keefe *et al.*, *Proc. Natl. Acad. Sci.* 93: 14670-675 (1996).

**[0161]** PNA nucleic acids can be used in prognostic, diagnostic, and therapeutic applications. For example, PNAs can be used as antisense or antigene agents for sequence-specific modulation of gene expression by, for example, inducing transcription or translation arrest or inhibiting replication. PNA nucleic acid molecules can also be used in the analysis of single base pair mutations in a gene, (*e.g.*, by PNA-directed PCR clamping); as "artificial restriction enzymes" when used in combination with other



enzymes, (e.g., S1 nucleases (Hyrup (1996) supra)); or as probes or primers for DNA sequencing or hybridization (Hyrup *et al.*, (1996) supra; Perry-O'Keefe supra).

[0162] In other embodiments, oligonucleotides may include other appended groups such as peptides (e.g., for targeting host cell receptors *in vivo*), or agents facilitating transport across cell membranes (see e.g., Letsinger *et al.*, Proc. Natl. Acad. Sci. USA 86: 6553-6556 (1989); Lemaitre *et al.*, Proc. Natl. Acad. Sci. USA 84: 648-652 (1987); PCT Publication No. W088/09810) or the blood-brain barrier (see, e.g., PCT Publication No. W089/10134). In addition, oligonucleotides can be modified with hybridization-triggered cleavage agents (See, e.g., Krol *et al.*, Bio-Techniques 6: 958-976 (1988)) or intercalating agents. (See, e.g., Zon, Pharm. Res. 5: 539-549 (1988) ). To this end, the oligonucleotide may be conjugated to another molecule, (e.g., a peptide, hybridization triggered cross-linking agent, transport agent, or hybridization-triggered cleavage agent).

[0163] Also included herein are molecular beacon oligonucleotide primer and probe molecules having one or more regions complementary to a nucleotide sequence of SEQ ID NO: 1-12 or a substantially identical sequence thereof, two complementary regions one having a fluorophore and one a quencher such that the molecular beacon is useful for quantifying the presence of the nucleic acid in a sample. Molecular beacon nucleic acids are described, for example, in Lizardi *et al.*, U.S. Patent No. 5,854,033; Nazarenko *et al.*, U.S. Patent No. 5,866,336, and Livak *et al.*, U.S. Patent 5,876,930.

#### Antibodies

[0164] The term "antibody" as used herein refers to an immunoglobulin molecule or immunologically active portion thereof, i.e., an antigen-binding portion. Examples of immunologically active portions of immunoglobulin molecules include F(ab) and F(ab')<sub>2</sub> fragments which can be generated by treating the antibody with an enzyme such as pepsin. An antibody sometimes is a polyclonal, monoclonal, recombinant (e.g., a chimeric or humanized), fully human, non-human (e.g., murine), or a single chain antibody. An antibody may have effector function and can fix complement, and is sometimes coupled to a toxin or imaging agent.

[0165] A full-length polypeptide or antigenic peptide fragment encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleotide sequence can be used as an immunogen or can be used to identify antibodies made with other immunogens, e.g., cells, membrane preparations, and the like. An antigenic peptide often includes at least 8 amino acid residues of the amino acid sequences encoded by a nucleotide sequence of SEQ ID NO: 1-12, or substantially identical sequence thereof, and encompasses an epitope. Antigenic peptides sometimes include 10 or more amino acids, 15 or more amino acids, 20 or more amino acids, or 30 or more amino acids. Hydrophilic and hydrophobic fragments of polypeptides sometimes are used as immunogens.

[0166] Epitopes encompassed by the antigenic peptide are regions located on the surface of the polypeptide (*e.g.*, hydrophilic regions) as well as regions with high antigenicity. For example, an Emini surface probability analysis of the human polypeptide sequence can be used to indicate the regions that have a particularly high probability of being localized to the surface of the polypeptide and are thus likely to constitute surface residues useful for targeting antibody production. The antibody may bind an epitope on any domain or region on polypeptides described herein.

[0167] Also, chimeric, humanized, and completely human antibodies are useful for applications which include repeated administration to subjects. Chimeric and humanized monoclonal antibodies, comprising both human and non-human portions, can be made using standard recombinant DNA techniques. Such chimeric and humanized monoclonal antibodies can be produced by recombinant DNA techniques known in the art, for example using methods described in Robinson *et al* International Application No. PCT/US86/02269; Akira, *et al* European Patent Application 184,187; Taniguchi, M., European Patent Application 171,496; Morrison *et al* European Patent Application 173,494; Neuberger *et al* PCT International Publication No. WO 86/01533; Cabilly *et al* U.S. Patent No. 4,816,567; Cabilly *et al* European Patent Application 125,023; Better *et al.*, Science 240: 1041-1043 (1988); Liu *et al.*, Proc. Natl. Acad. Sci. USA 84: 3439-3443 (1987); Liu *et al.*, J. Immunol. 139: 3521-3526 (1987); Sun *et al.*, Proc. Natl. Acad. Sci. USA 84: 214-218 (1987); Nishimura *et al.*, Canc. Res. 47: 999-1005 (1987); Wood *et al.*, Nature 314: 446-449 (1985); and Shaw *et al.*, J. Natl. Cancer Inst. 80: 1553-1559 (1988); Morrison, S. L., Science 229: 1202-1207 (1985); Oi *et al.*, BioTechniques 4: 214 (1986); Winter U.S. Patent 5,225,539; Jones *et al.*, Nature 321: 552-525 (1986); Verhoeyan *et al.*, Science 239: 1534; and Beidler *et al.*, J. Immunol. 141: 4053-4060 (1988).

[0168] Completely human antibodies are particularly desirable for therapeutic treatment of human patients. Such antibodies can be produced using transgenic mice that are incapable of expressing endogenous immunoglobulin heavy and light chains genes, but which can express human heavy and light chain genes. See, for example, Lonberg and Huszar, Int. Rev. Immunol. 13: 65-93 (1995); and U.S. Patent Nos. 5,625,126; 5,633,425; 5,569,825; 5,661,016; and 5,545,806. In addition, companies such as Abgenix, Inc. (Fremont, CA) and Medarex, Inc. (Princeton, NJ), can be engaged to provide human antibodies directed against a selected antigen using technology similar to that described above. Completely human antibodies that recognize a selected epitope also can be generated using a technique referred to as "guided selection." In this approach a selected non-human monoclonal antibody (*e.g.*, a murine antibody) is used to guide the selection of a completely human antibody recognizing the same epitope. This technology is described for example by Jespers *et al.*, Bio/Technology 12: 899-903 (1994).

[0169] Antibody can be a single chain antibody. A single chain antibody (scFV) can be engineered (see, *e.g.*, Colcher *et al.*, Ann. N Y Acad. Sci. 880: 263-80 (1999); and Reiter, Clin. Cancer Res. 2: 245-

52 (1996)). Single chain antibodies can be dimerized or multimerized to generate multivalent antibodies having specificities for different epitopes of the same target polypeptide.

[0170] Antibodies also may be selected or modified so that they exhibit reduced or no ability to bind an Fc receptor. For example, an antibody may be an isotype or subtype, fragment or other mutant, which does not support binding to an Fc receptor (*e.g.*, it has a mutagenized or deleted Fc receptor binding region).

[0171] Also, an antibody (or fragment thereof) may be conjugated to a therapeutic moiety such as a cytotoxin, a therapeutic agent or a radioactive metal ion. A cytotoxin or cytotoxic agent includes any agent that is detrimental to cells. Examples include taxol, cytochalasin B, gramicidin D, ethidium bromide, emetine, mitomycin, etoposide, tenoposide, vincristine, vinblastine, colchicin, doxorubicin, daunorubicin, dihydroxy anthracin dione, mitoxantrone, mithramycin, actinomycin D, 1 dehydrotestosterone, glucocorticoids, procaine, tetracaine, lidocaine, propranolol, and puromycin and analogs or homologs thereof. Therapeutic agents include, but are not limited to, antimetabolites (*e.g.*, methotrexate, 6-mercaptopurine, 6-thioguanine, cytarabine, 5-fluorouracil decarbazine), alkylating agents (*e.g.*, mechlorethamine, thiotepa chlorambucil, melphalan, carmustine (BCNU) and lomustine (CCNU), cyclophosphamide, busulfan, dibromomannitol, streptozotocin, mitomycin C, and cis-dichlorodiamine platinum (II) (DDP) cisplatin), anthracyclines (*e.g.*, daunorubicin (formerly daunomycin) and doxorubicin), antibiotics (*e.g.*, dactinomycin (formerly actinomycin), bleomycin, mithramycin, and anthramycin (AMC)), and anti-mitotic agents (*e.g.*, vincristine and vinblastine).

[0172] Antibody conjugates can be used for modifying a given biological response. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor,  $\gamma$ -interferon,  $\alpha$ -interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Also, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980, for example.

[0173] An antibody (*e.g.*, monoclonal antibody) can be used to isolate target polypeptides by standard techniques, such as affinity chromatography or immunoprecipitation. Moreover, an antibody can be used to detect a target polypeptide (*e.g.*, in a cellular lysate or cell supernatant) in order to evaluate the abundance and pattern of expression of the polypeptide. Antibodies can be used diagnostically to monitor polypeptide levels in tissue as part of a clinical testing procedure, *e.g.*, to determine the efficacy of a given treatment regimen. Detection can be facilitated by coupling (*i.e.*, physically linking) the

antibody to a detectable substance (i.e., antibody labeling). Examples of detectable substances include various enzymes, prosthetic groups, fluorescent materials, luminescent materials, bioluminescent materials, and radioactive materials. Examples of suitable enzymes include horseradish peroxidase, alkaline phosphatase,  $\beta$ -galactosidase, or acetylcholinesterase; examples of suitable prosthetic group complexes include streptavidin/biotin and avidin/biotin; examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin; an example of a luminescent material includes luminol; examples of bioluminescent materials include luciferase, luciferin, and aequorin, and examples of suitable radioactive material include  $^{125}\text{I}$ ,  $^{131}\text{I}$ ,  $^{35}\text{S}$  or  $^3\text{H}$ . Also, an antibody can be utilized as a test molecule for determining whether it can treat breast cancer, and as a therapeutic for administration to a subject for treating breast cancer.

[0174] An antibody can be made by immunizing with a purified antigen, or a fragment thereof, *e.g.*, a fragment described herein, a membrane associated antigen, tissues, *e.g.*, crude tissue preparations, whole cells, preferably living cells, lysed cells, or cell fractions.

[0175] Included herein are antibodies which bind only a native polypeptide, only denatured or otherwise non-native polypeptide, or which bind both, as well as those having linear or conformational epitopes. Conformational epitopes sometimes can be identified by selecting antibodies that bind to native but not denatured polypeptide. Also featured are antibodies that specifically bind to a polypeptide variant associated with breast cancer.

#### Screening Assays

[0176] Featured herein are methods for identifying a candidate therapeutic for treating breast cancer. The methods comprise contacting a test molecule with a target molecule in a system. A "target molecule" as used herein refers to a nucleic acid of SEQ ID NO: 1-12, a substantially identical nucleic acid thereof, or a fragment thereof, and an encoded polypeptide of the foregoing. The method also comprises determining the presence or absence of an interaction between the test molecule and the target molecule, where the presence of an interaction between the test molecule and the nucleic acid or polypeptide identifies the test molecule as a candidate breast cancer therapeutic. The interaction between the test molecule and the target molecule may be quantified.

[0177] Test molecules and candidate therapeutics include, but are not limited to, compounds, antisense nucleic acids, siRNA molecules, ribozymes, polypeptides or proteins encoded by a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acids, or a substantially identical sequence or fragment thereof, and immunotherapeutics (*e.g.*, antibodies and HLA-presented polypeptide fragments). A test molecule or candidate therapeutic may act as a modulator of target molecule concentration or target

molecule function in a system. A “modulator” may agonize (i.e., up-regulates) or antagonize (i.e., down-regulates) a target molecule concentration partially or completely in a system by affecting such cellular functions as DNA replication and/or DNA processing (e.g., DNA methylation or DNA repair), RNA transcription and/or RNA processing (e.g., removal of intronic sequences and/or translocation of spliced mRNA from the nucleus), polypeptide production (e.g., translation of the polypeptide from mRNA), and/or polypeptide post-translational modification (e.g., glycosylation, phosphorylation, and proteolysis of pro-polypeptides). A modulator may also agonize or antagonize a biological function of a target molecule partially or completely, where the function may include adopting a certain structural conformation, interacting with one or more binding partners, ligand binding, catalysis (e.g., phosphorylation, dephosphorylation, hydrolysis, methylation, and isomerization), and an effect upon a cellular event (e.g., effecting progression of breast cancer).

[0178] As used herein, the term “system” refers to a cell free *in vitro* environment and a cell-based environment such as a collection of cells, a tissue, an organ, or an organism. A system is “contacted” with a test molecule in a variety of manners, including adding molecules in solution and allowing them to interact with one another by diffusion, cell injection, and any administration routes in an animal. As used herein, the term “interaction” refers to an effect of a test molecule on test molecule, where the effect sometimes is binding between the test molecule and the target molecule, and sometimes is an observable change in cells, tissue, or organism.

[0179] There are many standard methods for detecting the presence or absence of an interaction between a test molecule and a target molecule. For example, titrametric, acidimetric, radiometric, NMR, monolayer, polarographic, spectrophotometric, fluorescent, and ESR assays probative of a target molecule interaction may be utilized.

[0180] In general, an interaction can be determined by labeling the test molecule and/or the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule, where the label is covalently or non-covalently attached to the test molecule or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule. The label is sometimes a radioactive molecule such as  $^{125}\text{I}$ ,  $^{131}\text{I}$ ,  $^{35}\text{S}$  or  $^3\text{H}$ , which can be detected by direct counting of radioemission or by scintillation counting. Also, enzymatic labels such as horseradish peroxidase, alkaline phosphatase, or luciferase may be utilized where the enzymatic label can be detected by determining conversion of an appropriate substrate to product. Also, presence or absence of an interaction can be determined without labeling. For example, a microphysiometer (e.g., Cytosensor) is an analytical instrument that measures the rate at which a cell acidifies its environment using a light-addressable potentiometric sensor (LAPS). Changes in this acidification rate can be used as an indication of an interaction between a test molecule and *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* (McConnell, H. M. et al., Science 257: 1906-1912 (1992)).

[0181] In cell-based systems, cells typically include a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid or polypeptide or variants thereof and are often of mammalian origin, although the cell can be of any origin. Whole cells, cell homogenates, and cell fractions (e.g., cell membrane fractions) can be subjected to analysis. Where interactions between a test molecule with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or variant thereof are monitored, soluble and/or membrane bound forms of the polypeptide or variant may be utilized. Where membrane-bound forms of the polypeptide are used, it may be desirable to utilize a solubilizing agent. Examples of such solubilizing agents include non-ionic detergents such as n-octylglucoside, n-dodecylglucoside, n-dodecylmaltoside, octanoyl-N-methylglucamide, decanoyl-N-methylglucamide, Triton® X-100, Triton® X-114, Thesit®, Isotridecypoly(ethylene glycol ether)n, 3-[(3-cholamidopropyl)dimethylamminio]-1-propane sulfonate (CHAPS), 3-[(3-cholamidopropyl)dimethylamminio]-2-hydroxy-1-propane sulfonate (CHAPSO), or N-dodecyl-N,N-dimethyl-3-ammonio-1-propane sulfonate.

[0182] An interaction between two molecules also can be detected by monitoring fluorescence energy transfer (FET) (see, for example, Lakowicz et al., U.S. Patent No. 5,631,169; Stavrianopoulos et al. U.S. Patent No. 4,868,103). A fluorophore label on a first, “donor” molecule is selected such that its emitted fluorescent energy will be absorbed by a fluorescent label on a second, “acceptor” molecule, which in turn is able to fluoresce due to the absorbed energy. Alternately, the “donor” polypeptide molecule may simply utilize the natural fluorescent energy of tryptophan residues. Labels are chosen that emit different wavelengths of light, such that the “acceptor” molecule label may be differentiated from that of the “donor”. Since the efficiency of energy transfer between the labels is related to the distance separating the molecules, the spatial relationship between the molecules can be assessed. In a situation in which binding occurs between the molecules, the fluorescent emission of the “acceptor” molecule label in the assay should be maximal. An FET binding event can be conveniently measured through standard fluorometric detection means well known in the art (e.g., using a fluorimeter).

[0183] In another embodiment, determining the presence or absence of an interaction between a test molecule and a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule can be effected by using real-time Biomolecular Interaction Analysis (BIA) (see, e.g., Sjolander & Urbanicz, Anal. Chem. 63: 2338-2345 (1991) and Szabo et al., Curr. Opin. Struct. Biol. 5: 699-705 (1995)). “Surface plasmon resonance” or “BIA” detects biospecific interactions in real time, without labeling any of the interactants (e.g., BIAcore). Changes in the mass at the binding surface (indicative of a binding event) result in alterations of the refractive index of light near the surface (the optical phenomenon of surface plasmon resonance (SPR)), resulting in a detectable signal which can be used as an indication of real-time reactions between biological molecules.

[0184] In another embodiment, the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule or test molecules are anchored to a solid phase. The *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule/test molecule complexes anchored to the solid phase can be detected at the end of the reaction. The target *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule is often anchored to a solid surface, and the test molecule, which is not anchored, can be labeled, either directly or indirectly, with detectable labels discussed herein.

[0185] It may be desirable to immobilize a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule, an anti-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antibody, or test molecules to facilitate separation of complexed from uncomplexed forms of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules and test molecules, as well as to accommodate automation of the assay. Binding of a test molecule to a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule can be accomplished in any vessel suitable for containing the reactants. Examples of such vessels include microtiter plates, test tubes, and micro-centrifuge tubes. In one embodiment, a fusion polypeptide can be provided which adds a domain that allows a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule to be bound to a matrix. For example, glutathione-S-transferase/*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* fusion polypeptides or glutathione-S-transferase/target fusion polypeptides can be adsorbed onto glutathione sepharose beads (Sigma Chemical, St. Louis, MO) or glutathione derivitized microtiter plates, which are then combined with the test compound or the test compound and either the non-adsorbed target polypeptide or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, and the mixture incubated under conditions conducive to complex formation (e.g., at physiological conditions for salt and pH). Following incubation, the beads or microtiter plate wells are washed to remove any unbound components, the matrix immobilized in the case of beads, complex determined either directly or indirectly, for example, as described above. Alternatively, the complexes can be dissociated from the matrix, and the level of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* binding or activity determined using standard techniques.

[0186] Other techniques for immobilizing a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule on matrices include using biotin and streptavidin. For example, biotinylated *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or target molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, IL), and immobilized in the wells of streptavidin-coated 96 well plates (Pierce Chemical).

[0187] In order to conduct the assay, the non-immobilized component is added to the coated surface containing the anchored component. After the reaction is complete, unreacted components are removed (e.g., by washing) under conditions such that any complexes formed will remain immobilized on the solid surface. The detection of complexes anchored on the solid surface can be accomplished in a

number of ways. Where the previously non-immobilized component is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the previously non-immobilized component is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the immobilized component (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody).

[0188] In one embodiment, this assay is performed utilizing antibodies reactive with *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or test molecules but which do not interfere with binding of the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide to its test molecule. Such antibodies can be derivitized to the wells of the plate, and unbound target or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide trapped in the wells by antibody conjugation. Methods for detecting such complexes, in addition to those described above for the GST-immobilized complexes, include immunodetection of complexes using antibodies reactive with the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or target molecule, as well as enzyme-linked assays which rely on detecting an enzymatic activity associated with the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or test molecule.

[0189] Alternatively, cell free assays can be conducted in a liquid phase. In such an assay, the reaction products are separated from unreacted components, by any of a number of standard techniques, including but not limited to: differential centrifugation (see, for example, Rivas, G., and Minton, A. P., Trends Biochem Sci Aug;18(8): 284-7 (1993)); chromatography (gel filtration chromatography, ion-exchange chromatography); electrophoresis (see, e.g., Ausubel et al., eds. Current Protocols in Molecular Biology, J. Wiley: New York (1999)); and immunoprecipitation (see, for example, Ausubel, F. et al., eds. Current Protocols in Molecular Biology, J. Wiley: New York (1999)). Such resins and chromatographic techniques are known to one skilled in the art (see, e.g., Heegaard, J Mol. Recognit. Winter; 11(1-6): 141-8 (1998); Hage & Tweed, J. Chromatogr. B Biomed. Sci. Appl. Oct 10; 699 (1-2): 499-525 (1997)). Further, fluorescence energy transfer may also be conveniently utilized, as described herein, to detect binding without further purification of the complex from solution.

[0190] In another embodiment, modulators of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression are identified. For example, a cell or cell free mixture is contacted with a candidate compound and the expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide evaluated relative to the level of expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide in the absence of the candidate compound. When expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide is greater in the presence of the candidate compound than in its absence, the candidate compound is identified as a stimulator of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide expression. Alternatively, when expression of



*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide is less (statistically significantly less) in the presence of the candidate compound than in its absence, the candidate compound is identified as an inhibitor of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide expression. The level of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide expression can be determined by methods described herein for detecting *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* mRNA or polypeptide.

[0191] In another embodiment, binding partners that interact with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule are detected. The *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules can interact with one or more cellular or extracellular macromolecules, such as polypeptides, in vivo, and these molecules that interact with *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules are referred to herein as “binding partners.” Molecules that disrupt such interactions can be useful in regulating the activity of the target gene product. Such molecules can include, but are not limited to molecules such as antibodies, peptides, and small molecules. Target genes/products for use in this embodiment often are the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* genes herein identified. In an alternative embodiment, provided is a method for determining the ability of the test compound to modulate the activity of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide through modulation of the activity of a downstream effector of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* target molecule. For example, the activity of the effector molecule on an appropriate target can be determined, or the binding of the effector to an appropriate target can be determined, as previously described.

[0192] To identify compounds that interfere with the interaction between the target gene product and its cellular or extracellular binding partner(s), e.g., a substrate, a reaction mixture containing the target gene product and the binding partner is prepared, under conditions and for a time sufficient, to allow the two products to form complex. In order to test an inhibitory agent, the reaction mixture is provided in the presence and absence of the test compound. The test compound can be initially included in the reaction mixture, or can be added at a time subsequent to the addition of the target gene and its cellular or extracellular binding partner. Control reaction mixtures are incubated without the test compound or with a placebo. The formation of any complexes between the target gene product and the cellular or extracellular binding partner is then detected. The formation of a complex in the control reaction, but not in the reaction mixture containing the test compound, indicates that the compound interferes with the interaction of the target gene product and the interactive binding partner. Additionally, complex formation within reaction mixtures containing the test compound and normal target gene product can also be compared to complex formation within reaction mixtures containing the test compound and mutant target gene product. This comparison can be important in those cases where it is desirable to identify compounds that disrupt interactions of mutant but not normal target gene products.

[0193] These assays can be conducted in a heterogeneous or homogeneous format. Heterogeneous assays involve anchoring either the target gene product or the binding partner onto a solid phase, and detecting complexes anchored on the solid phase at the end of the reaction. In homogeneous assays, the entire reaction is carried out in a liquid phase. In either approach, the order of addition of reactants can be varied to obtain different information about the compounds being tested. For example, test compounds that interfere with the interaction between the target gene products and the binding partners, e.g., by competition, can be identified by conducting the reaction in the presence of the test substance. Alternatively, test compounds that disrupt preformed complexes, e.g., compounds with higher binding constants that displace one of the components from the complex, can be tested by adding the test compound to the reaction mixture after complexes have been formed. The various formats are briefly described below.

[0194] In a heterogeneous assay system, either the target gene product or the interactive cellular or extracellular binding partner, is anchored onto a solid surface (e.g., a microtiter plate), while the non-anchored species is labeled, either directly or indirectly. The anchored species can be immobilized by non-covalent or covalent attachments. Alternatively, an immobilized antibody specific for the species to be anchored can be used to anchor the species to the solid surface.

[0195] In order to conduct the assay, the partner of the immobilized species is exposed to the coated surface with or without the test compound. After the reaction is complete, unreacted components are removed (e.g., by washing) and any complexes formed will remain immobilized on the solid surface. Where the non-immobilized species is pre-labeled, the detection of label immobilized on the surface indicates that complexes were formed. Where the non-immobilized species is not pre-labeled, an indirect label can be used to detect complexes anchored on the surface; e.g., using a labeled antibody specific for the initially non-immobilized species (the antibody, in turn, can be directly labeled or indirectly labeled with, e.g., a labeled anti-Ig antibody). Depending upon the order of addition of reaction components, test compounds that inhibit complex formation or that disrupt preformed complexes can be detected.

[0196] Alternatively, the reaction can be conducted in a liquid phase in the presence or absence of the test compound, the reaction products separated from unreacted components, and complexes detected; e.g., using an immobilized antibody specific for one of the binding components to anchor any complexes formed in solution, and a labeled antibody specific for the other partner to detect anchored complexes. Again, depending upon the order of addition of reactants to the liquid phase, test compounds that inhibit complex or that disrupt preformed complexes can be identified.

[0197] In an alternate embodiment, a homogeneous assay can be used. For example, a preformed complex of the target gene product and the interactive cellular or extracellular binding partner product is prepared in that either the target gene products or their binding partners are labeled, but the signal

generated by the label is quenched due to complex formation (see, e.g., U.S. Patent No. 4,109,496 that utilizes this approach for immunoassays). The addition of a test substance that competes with and displaces one of the species from the preformed complex will result in the generation of a signal above background. In this way, test substances that disrupt target gene product-binding partner interaction can be identified.

[0198] Also, binding partners of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules can be identified in a two-hybrid assay or three-hybrid assay (see, e.g., U.S. Patent No. 5,283,317; Zervos et al., Cell 72:223-232 (1993); Madura et al., J. Biol. Chem. 268: 12046-12054 (1993); Bartel et al., Biotechniques 14: 920-924 (1993); Iwabuchi et al., Oncogene 8: 1693-1696 (1993); and Brent WO94/10300), to identify other polypeptides, which bind to or interact with *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* (“*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*-binding polypeptides” or “*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*-bp”) and are involved in *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity. Such *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*-bps can be activators or inhibitors of signals by the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptides or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* targets as, for example, downstream elements of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*-mediated signaling pathway.

[0199] A two-hybrid system is based on the modular nature of most transcription factors, which consist of separable DNA-binding and activation domains. Briefly, the assay utilizes two different DNA constructs. In one construct, the gene that codes for a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide is fused to a gene encoding the DNA binding domain of a known transcription factor (e.g., GAL-4). In the other construct, a DNA sequence, from a library of DNA sequences, that encodes an unidentified polypeptide (“prey” or “sample”) is fused to a gene that codes for the activation domain of the known transcription factor. (Alternatively the: *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide can be the fused to the activator domain.) If the “bait” and the “prey” polypeptides are able to interact, in vivo, forming a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*-dependent complex, the DNA-binding and activation domains of the transcription factor are brought into close proximity. This proximity allows transcription of a reporter gene (e.g., LacZ) which is operably linked to a transcriptional regulatory site responsive to the transcription factor. Expression of the reporter gene can be detected and cell colonies containing the functional transcription factor can be isolated and used to obtain the cloned gene which encodes the polypeptide which interacts with the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide.

[0200] Candidate therapeutics for treating breast cancer are identified from a group of test molecules that interact with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid or polypeptide. Test molecules are normally ranked according to the degree with which they interact or modulate (e.g.,

agonize or antagonize) DNA replication and/or processing, RNA transcription and/or processing, polypeptide production and/or processing, and/or function of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules, for example, and then top ranking modulators are selected. In a preferred embodiment, the candidate therapeutic (i.e., test molecule) acts as a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antagonist. Also, pharmacogenomic information described herein can determine the rank of a modulator. Candidate therapeutics typically are formulated for administration to a subject.

#### Therapeutic Treatments

[0201] Formulations or pharmaceutical compositions typically include in combination with a pharmaceutically acceptable carrier, a compound, an antisense nucleic acid, a ribozyme, an antibody, a binding partner that interacts with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid, or a fragment thereof. The formulated molecule may be one that is identified by a screening method described above. As used herein, the term “pharmaceutically acceptable carrier” includes solvents, dispersion media, coatings, antibacterial and antifungal agents, isotonic and absorption delaying agents, and the like, compatible with pharmaceutical administration. Supplementary active compounds can also be incorporated into the compositions.

[0202] A pharmaceutical composition is formulated to be compatible with its intended route of administration. Examples of routes of administration include parenteral, e.g., intravenous, intradermal, subcutaneous, oral (e.g., inhalation), transdermal (topical), transmucosal, and rectal administration. Solutions or suspensions used for parenteral, intradermal, or subcutaneous application can include the following components: a sterile diluent such as water for injection, saline solution, fixed oils, polyethylene glycols, glycerin, propylene glycol or other synthetic solvents; antibacterial agents such as benzyl alcohol or methyl parabens; antioxidants such as ascorbic acid or sodium bisulfite; chelating agents such as ethylenediaminetetraacetic acid; buffers such as acetates, citrates or phosphates and agents for the adjustment of tonicity such as sodium chloride or dextrose. pH can be adjusted with acids or bases, such as hydrochloric acid or sodium hydroxide. The parenteral preparation can be enclosed in ampoules, disposable syringes or multiple dose vials made of glass or plastic.

[0203] Oral compositions generally include an inert diluent or an edible carrier. For the purpose of oral therapeutic administration, the active compound can be incorporated with excipients and used in the form of tablets, troches, or capsules, e.g., gelatin capsules. Oral compositions can also be prepared using a fluid carrier for use as a mouthwash. Pharmaceutically compatible binding agents, and/or adjuvant materials can be included as part of the composition. The tablets, pills, capsules, troches and the like can contain any of the following ingredients, or compounds of a similar nature: a binder such as

microcrystalline cellulose, gum tragacanth or gelatin; an excipient such as starch or lactose, a disintegrating agent such as alginic acid, Primogel, or corn starch; a lubricant such as magnesium stearate or Sterotes; a glidant such as colloidal silicon dioxide; a sweetening agent such as sucrose or saccharin; or a flavoring agent such as peppermint, methyl salicylate, or orange flavoring.

[0204] Pharmaceutical compositions suitable for injectable use include sterile aqueous solutions (where water soluble) or dispersions and sterile powders for the extemporaneous preparation of sterile injectable solutions or dispersion. For intravenous administration, suitable carriers include physiological saline, bacteriostatic water, Cremophor EL™ (BASF, Parsippany, NJ) or phosphate buffered saline (PBS). In all cases, the composition must be sterile and should be fluid to the extent that easy syringability exists. It should be stable under the conditions of manufacture and storage and must be preserved against the contaminating action of microorganisms such as bacteria and fungi. The carrier can be a solvent or dispersion medium containing, for example, water, ethanol, polyol (for example, glycerol, propylene glycol, and liquid polyethylene glycol, and the like), and suitable mixtures thereof. The proper fluidity can be maintained, for example, by the use of a coating such as lecithin, by the maintenance of the required particle size in the case of dispersion and by the use of surfactants. Prevention of the action of microorganisms can be achieved by various antibacterial and antifungal agents, for example, parabens, chlorobutanol, phenol, ascorbic acid, thimerosal, and the like. In many cases, isotonic agents, for example, sugars, polyalcohols such as mannitol, sorbitol, sodium chloride sometimes are included in the composition. Prolonged absorption of the injectable compositions can be brought about by including in the composition an agent which delays absorption, for example, aluminum monostearate and gelatin.

[0205] Sterile injectable solutions can be prepared by incorporating the active compound in the required amount in an appropriate solvent with one or a combination of ingredients enumerated above, as required, followed by filtered sterilization. Generally, dispersions are prepared by incorporating the active compound into a sterile vehicle which contains a basic dispersion medium and the required other ingredients from those enumerated above. In the case of sterile powders for the preparation of sterile injectable solutions, methods of preparation often utilized are vacuum drying and freeze-drying which yields a powder of the active ingredient plus any additional desired ingredient from a previously sterile-filtered solution thereof.

[0206] For administration by inhalation, the compounds are delivered in the form of an aerosol spray from pressured container or dispenser which contains a suitable propellant, e.g., a gas such as carbon dioxide, or a nebulizer.

[0207] Systemic administration can also be by transmucosal or transdermal means. For transmucosal or transdermal administration, penetrants appropriate to the barrier to be permeated are used in the formulation. Such penetrants are generally known in the art, and include, for example, for

transmucosal administration, detergents, bile salts, and fusidic acid derivatives. Transmucosal administration can be accomplished through the use of nasal sprays or suppositories. For transdermal administration, the active compounds are formulated into ointments, salves, gels, or creams as generally known in the art. Molecules can also be prepared in the form of suppositories (e.g., with conventional suppository bases such as cocoa butter and other glycerides) or retention enemas for rectal delivery.

[0208] In one embodiment, active molecules are prepared with carriers that will protect the compound against rapid elimination from the body, such as a controlled release formulation, including implants and microencapsulated delivery systems. Biodegradable, biocompatible polymers can be used, such as ethylene vinyl acetate, polyanhydrides, polyglycolic acid, collagen, polyorthoesters, and polylactic acid. Methods for preparation of such formulations will be apparent to those skilled in the art. Materials can also be obtained commercially from Alza Corporation and Nova Pharmaceuticals, Inc. Liposomal suspensions (including liposomes targeted to infected cells with monoclonal antibodies to viral antigens) can also be used as pharmaceutically acceptable carriers. These can be prepared according to methods known to those skilled in the art, for example, as described in U.S. Patent No. 4,522,811.

[0209] It is advantageous to formulate oral or parenteral compositions in dosage unit form for ease of administration and uniformity of dosage. Dosage unit form as used herein refers to physically discrete units suited as unitary dosages for the subject to be treated; each unit containing a predetermined quantity of active compound calculated to produce the desired therapeutic effect in association with the required pharmaceutical carrier.

[0210] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD<sub>50</sub> (the dose lethal to 50% of the population) and the ED<sub>50</sub> (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD<sub>50</sub>/ED<sub>50</sub>. Molecules which exhibit high therapeutic indices often are utilized. While molecules that exhibit toxic side effects may be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0211] The data obtained from the cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such molecules often lies within a range of circulating concentrations that include the ED<sub>50</sub> with little or no toxicity. The dosage may vary within this range depending upon the dosage form employed and the route of administration utilized. For any molecules used in the methods described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose may be formulated in animal models to achieve a circulating

plasma concentration range that includes the  $IC_{50}$  (i.e., the concentration of the test compound which achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma may be measured, for example, by high performance liquid chromatography.

[0212] As defined herein, a therapeutically effective amount of protein or polypeptide (i.e., an effective dosage) ranges from about 0.001 to 30 mg/kg body weight, sometimes about 0.01 to 25 mg/kg body weight, often about 0.1 to 20 mg/kg body weight, and more often about 1 to 10 mg/kg, 2 to 9 mg/kg, 3 to 8 mg/kg, 4 to 7 mg/kg, or 5 to 6 mg/kg body weight. The protein or polypeptide can be administered one time per week for between about 1 to 10 weeks, sometimes between 2 to 8 weeks, often between about 3 to 7 weeks, and more often for about 4, 5, or 6 weeks. The skilled artisan will appreciate that certain factors may influence the dosage and timing required to effectively treat a subject, including but not limited to the severity of the disease or disorder, previous treatments, the general health and/or age of the subject, and other diseases present. Moreover, treatment of a subject with a therapeutically effective amount of a protein, polypeptide, or antibody can include a single treatment, or sometimes can include a series of treatments.

[0213] With regard to polypeptide formulations, featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide, where the subject comprises a second *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide having one or more polymorphic variations associated with cancer, and where the first polypeptide comprises fewer polymorphic variations associated with cancer than the second polypeptide. The first and second polypeptides are encoded by a nucleic acid which comprises a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-12; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-12; a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence of SEQ ID NO: 1-12 and a nucleotide sequence 90% or more identical to a nucleotide sequence of SEQ ID NO: 1-12. The subject is often a human.

[0214] For antibodies, a dosage of 0.1 mg/kg of body weight (generally 10 mg/kg to 20 mg/kg) is often utilized. If the antibody is to act in the brain, a dosage of 50 mg/kg to 100 mg/kg is often appropriate. Generally, partially human antibodies and fully human antibodies have a longer half-life within the human body than other antibodies. Accordingly, lower dosages and less frequent administration is often possible. Modifications such as lipidation can be used to stabilize antibodies and to enhance uptake and tissue penetration (e.g., into the brain). A method for lipidation of antibodies is

described by Cruikshank et al., J. Acquired Immune Deficiency Syndromes and Human Retrovirology 14:193 (1997).

[0215] Antibody conjugates can be used for modifying a given biological response, the drug moiety is not to be construed as limited to classical chemical therapeutic agents. For example, the drug moiety may be a protein or polypeptide possessing a desired biological activity. Such proteins may include, for example, a toxin such as abrin, ricin A, pseudomonas exotoxin, or diphtheria toxin; a polypeptide such as tumor necrosis factor, .alpha.-interferon, .beta.-interferon, nerve growth factor, platelet derived growth factor, tissue plasminogen activator; or, biological response modifiers such as, for example, lymphokines, interleukin-1 ("IL-1"), interleukin-2 ("IL-2"), interleukin-6 ("IL-6"), granulocyte macrophage colony stimulating factor ("GM-CSF"), granulocyte colony stimulating factor ("G-CSF"), or other growth factors. Alternatively, an antibody can be conjugated to a second antibody to form an antibody heteroconjugate as described by Segal in U.S. Patent No. 4,676,980.

[0216] For compounds, exemplary doses include milligram or microgram amounts of the compound per kilogram of subject or sample weight, for example, about 1 microgram per kilogram to about 500 milligrams per kilogram, about 100 micrograms per kilogram to about 5 milligrams per kilogram, or about 1 microgram per kilogram to about 50 micrograms per kilogram. It is understood that appropriate doses of a small molecule depend upon the potency of the small molecule with respect to the expression or activity to be modulated. When one or more of these small molecules is to be administered to an animal (e.g., a human) in order to modulate expression or activity of a polypeptide or nucleic acid described herein, a physician, veterinarian, or researcher may, for example, prescribe a relatively low dose at first, subsequently increasing the dose until an appropriate response is obtained. In addition, it is understood that the specific dose level for any particular animal subject will depend upon a variety of factors including the activity of the specific compound employed, the age, body weight, general health, gender, and diet of the subject, the time of administration, the route of administration, the rate of excretion, any drug combination, and the degree of expression or activity to be modulated.

[0217] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid molecules can be inserted into vectors and used in gene therapy methods for treating breast cancer. Featured herein is a method for treating breast cancer in a subject, which comprises contacting one or more cells in the subject with a first *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid, where genomic DNA in the subject comprises a second *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid comprising one or more polymorphic variations associated with breast cancer, and where the first nucleic acid comprises fewer polymorphic variations associated with breast cancer. The first and second nucleic acids typically comprise a nucleotide sequence selected from the group consisting of the nucleotide sequence of SEQ ID NO: 1-5; a nucleotide sequence which encodes a polypeptide consisting of an amino acid sequence



encoded by a nucleotide sequence in SEQ ID NO: 1-5; a nucleotide sequence that is 90% or more identical to the nucleotide sequence of SEQ ID NO: 1-5, and a nucleotide sequence which encodes a polypeptide that is 90% or more identical to an amino acid sequence encoded by a nucleotide sequence in SEQ ID NO: 1-5. The subject often is a human.

[0218] Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Patent 5,328,470) or by stereotactic injection (see e.g., Chen et al., (1994) Proc. Natl. Acad. Sci. USA 91:3054-3057). Pharmaceutical preparations of gene therapy vectors can include a gene therapy vector in an acceptable diluent, or can comprise a slow release matrix in which the gene delivery vehicle is imbedded. Alternatively, where the complete gene delivery vector can be produced intact from recombinant cells (e.g., retroviral vectors) the pharmaceutical preparation can include one or more cells which produce the gene delivery system. Examples of gene delivery vectors are described herein.

[0219] Pharmaceutical compositions can be included in a container, pack, or dispenser together with instructions for administration.

[0220] Pharmaceutical compositions of active ingredients can be administered by any of the paths described herein for therapeutic and prophylactic methods for treating breast cancer. With regard to both prophylactic and therapeutic methods of treatment, such treatments may be specifically tailored or modified, based on knowledge obtained from pharmacogenomic analyses described herein. As used herein, the term "treatment" is defined as the application or administration of a therapeutic agent to a patient, or application or administration of a therapeutic agent to an isolated tissue or cell line from a patient, who has a disease, a symptom of disease or a predisposition toward a disease, with the purpose to cure, heal, alleviate, relieve, alter, remedy, ameliorate, improve or affect the disease, the symptoms of disease or the predisposition toward disease. A therapeutic agent includes, but is not limited to, small molecules, peptides, antibodies, ribozymes and antisense oligonucleotides.

[0221] Administration of a prophylactic agent can occur prior to the manifestation of symptoms characteristic of the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* aberrance, such that a disease or disorder is prevented or, alternatively, delayed in its progression. Depending on the type of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* aberrance, for example, a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecule, *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* agonist, or *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antagonist agent can be used for treating the subject. The appropriate agent can be determined based on screening assays described herein.

[0222] As discussed, successful treatment of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* disorders can be brought about by techniques that serve to inhibit the expression or activity of target gene products. For example, compounds (e.g., an agent identified using an assays described above) that

exhibit negative modulatory activity can be used to prevent and/or treat breast cancer. Such molecules can include, but are not limited to peptides, phosphopeptides, small organic or inorganic molecules, or antibodies (including, for example, polyclonal, monoclonal, humanized, anti-idiotypic, chimeric or single chain antibodies, and FAb, F(ab')<sub>2</sub> and FAb expression library fragments, scFV molecules, and epitope-binding fragments thereof).

[0223] Further, antisense and ribozyme molecules that inhibit expression of the target gene can also be used to reduce the level of target gene expression, thus effectively reducing the level of target gene activity. Still further, triple helix molecules can be utilized in reducing the level of target gene activity. Antisense, ribozyme and triple helix molecules are discussed above.

[0224] It is possible that the use of antisense, ribozyme, and/or triple helix molecules to reduce or inhibit mutant gene expression can also reduce or inhibit the transcription (triple helix) and/or translation (antisense, ribozyme) of mRNA produced by normal target gene alleles, such that the concentration of normal target gene product present can be lower than is necessary for a normal phenotype. In such cases, nucleic acid molecules that encode and express target gene polypeptides exhibiting normal target gene activity can be introduced into cells via gene therapy method. Alternatively, in instances where the target gene encodes an extracellular polypeptide, normal target gene polypeptide often is co-administered into the cell or tissue to maintain the requisite level of cellular or tissue target gene activity.

[0225] Another method by which nucleic acid molecules may be utilized in treating or preventing a disease characterized by *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression is through the use of aptamer molecules specific for *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Aptamers are nucleic acid molecules having a tertiary structure which permits them to specifically bind to polypeptide ligands (see, e.g., Osborne, et al., Curr. Opin. Chem. Biol.1(1): 5-9 (1997); and Patel, D. J., Curr. Opin. Chem. Biol. Jun;1(1): 32-46 (1997)). Since nucleic acid molecules may in many cases be more conveniently introduced into target cells than therapeutic polypeptide molecules may be, aptamers offer a method by which *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide activity may be specifically decreased without the introduction of drugs or other molecules which may have pluripotent effects.

[0226] Antibodies can be generated that are both specific for target gene product and that reduce target gene product activity. Such antibodies may, therefore, be administered in instances whereby negative modulatory techniques are appropriate for the treatment of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* disorders. For a description of antibodies, see the Antibody section above.

[0227] In circumstances where injection of an animal or a human subject with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or epitope for stimulating antibody production is harmful to the subject, it is possible to generate an immune response against *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1*

or *GALE* through the use of anti-idiotypic antibodies (see, for example, Herlyn, D., Ann. Med.;31(1): 66-78 (1999); and Bhattacharya-Chatterjee & Foon, Cancer Treat. Res.; 94: 51-68 (1998)). If an anti-idiotypic antibody is introduced into a mammal or human subject, it should stimulate the production of anti-anti-idiotypic antibodies, which should be specific to the *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide. Vaccines directed to a disease characterized by *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression may also be generated in this fashion.

[0228] In instances where the target antigen is intracellular and whole antibodies are used, internalizing antibodies may be utilized. Lipofectin or liposomes can be used to deliver the antibody or a fragment of the Fab region that binds to the target antigen into cells. Where fragments of the antibody are used, the smallest inhibitory fragment that binds to the target antigen often is utilized. For example, peptides having an amino acid sequence corresponding to the Fv region of the antibody can be used. Alternatively, single chain neutralizing antibodies that bind to intracellular target antigens can also be administered. Such single chain antibodies can be administered, for example, by expressing nucleotide sequences encoding single-chain antibodies within the target cell population (see e.g., Marasco et al., Proc. Natl. Acad. Sci. USA 90: 7889-7893 (1993)).

[0229] *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* molecules and compounds that inhibit target gene expression, synthesis and/or activity can be administered to a patient at therapeutically effective doses to prevent, treat or ameliorate *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* disorders. A therapeutically effective dose refers to that amount of the compound sufficient to result in amelioration of symptoms of the disorders.

[0230] Toxicity and therapeutic efficacy of such compounds can be determined by standard pharmaceutical procedures in cell cultures or experimental animals, e.g., for determining the LD<sub>50</sub> (the dose lethal to 50% of the population) and the ED<sub>50</sub> (the dose therapeutically effective in 50% of the population). The dose ratio between toxic and therapeutic effects is the therapeutic index and it can be expressed as the ratio LD<sub>50</sub>/ED<sub>50</sub>. Compounds that exhibit large therapeutic indices often are utilized. While compounds that exhibit toxic side effects can be used, care should be taken to design a delivery system that targets such compounds to the site of affected tissue in order to minimize potential damage to uninfected cells and, thereby, reduce side effects.

[0231] Data obtained from cell culture assays and animal studies can be used in formulating a range of dosage for use in humans. The dosage of such compounds often lies within a range of circulating concentrations that include the ED<sub>50</sub> with little or no toxicity. The dosage can vary within this range depending upon the dosage form employed and the route of administration utilized. For any compound used in a method described herein, the therapeutically effective dose can be estimated initially from cell culture assays. A dose can be formulated in animal models to achieve a circulating plasma concentration

range that includes the  $IC_{50}$  (i.e., the concentration of the test compound that achieves a half-maximal inhibition of symptoms) as determined in cell culture. Such information can be used to more accurately determine useful doses in humans. Levels in plasma can be measured, for example, by high performance liquid chromatography.

[0232] Another example of effective dose determination for an individual is the ability to directly assay levels of “free” and “bound” compound in the serum of the test subject. Such assays may utilize antibody mimics and/or “biosensors” that have been created through molecular imprinting techniques. The compound which is able to modulate *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is used as a template, or “imprinting molecule”, to spatially organize polymerizable monomers prior to their polymerization with catalytic reagents. The subsequent removal of the imprinted molecule leaves a polymer matrix which contains a repeated “negative image” of the compound and is able to selectively rebind the molecule under biological assay conditions. A detailed review of this technique can be seen in Ansell et al., *Current Opinion in Biotechnology* 7: 89-94 (1996) and in Shea, *Trends in Polymer Science* 2: 166-173 (1994). Such “imprinted” affinity matrixes are amenable to ligand-binding assays, whereby the immobilized monoclonal antibody component is replaced by an appropriately imprinted matrix. An example of the use of such matrixes in this way can be seen in Vlatakis, et al., *Nature* 361: 645-647 (1993). Through the use of isotope-labeling, the “free” concentration of compound which modulates the expression or activity of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* can be readily monitored and used in calculations of  $IC_{50}$ . Such “imprinted” affinity matrixes can also be designed to include fluorescent groups whose photon-emitting properties measurably change upon local and selective binding of target compound. These changes can be readily assayed in real time using appropriate fiberoptic devices, in turn allowing the dose in a test subject to be quickly optimized based on its individual  $IC_{50}$ . A rudimentary example of such a “biosensor” is discussed in Kriz et al., *Analytical Chemistry* 67: 2142-2144 (1995).

[0233] Provided herein are methods of modulating *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression or activity for therapeutic purposes. Accordingly, in an exemplary embodiment, the modulatory method involves contacting a cell with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* or agent that modulates one or more of the activities of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide activity associated with the cell. An agent that modulates *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide activity can be an agent as described herein, such as a nucleic acid or a polypeptide, a naturally-occurring target molecule of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide (e.g., a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* substrate or receptor), a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antibody, a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*

agonist or antagonist, a peptidomimetic of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* agonist or antagonist, or other small molecule.

[0234] In one embodiment, the agent stimulates one or more *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activities. Examples of such stimulatory agents include active *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide and a nucleic acid molecule encoding *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*. In another embodiment, the agent inhibits one or more *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activities. Examples of such inhibitory agents include antisense *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* nucleic acid molecules, anti-*ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* antibodies, and *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* inhibitors, and competitive inhibitors that target *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*. These modulatory methods can be performed in vitro (e.g., by culturing the cell with the agent) or, alternatively, in vivo (e.g., by administering the agent to a subject). As such, provided are methods of treating an individual afflicted with a disease or disorder characterized by aberrant or unwanted expression or activity of a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or nucleic acid molecule. In one embodiment, the method involves administering an agent (e.g., an agent identified by a screening assay described herein), or combination of agents that modulates (e.g., upregulates or downregulates) *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression or activity. In another embodiment, the method involves administering a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polypeptide or nucleic acid molecule as therapy to compensate for reduced, aberrant, or unwanted *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* expression or activity.

[0235] Stimulation of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is desirable in situations in which *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* is abnormally downregulated and/or in which increased *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is likely to have a beneficial effect. For example, stimulation of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is desirable in situations in which a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* is downregulated and/or in which increased *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is likely to have a beneficial effect. Likewise, inhibition of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is desirable in situations in which *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* is abnormally upregulated and/or in which decreased *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* activity is likely to have a beneficial effect.

#### Methods of Treatment

[0236] In another aspect, provided are methods for identifying a risk of cancer in an individual as described herein and, if a genetic predisposition is identified, treating that individual to delay or reduce or prevent the development of cancer. Such a procedure can be used to treat breast cancer. Optionally,

treating an individual for cancer may include inhibiting cellular proliferation, inhibiting metastasis, inhibiting invasion, or preventing tumor formation or growth as defined herein. Suitable treatments to prevent or reduce or delay breast cancer focus on inhibiting additional cellular proliferation, inhibiting metastasis, inhibiting invasion, and preventing further tumor formation or growth. Treatment usually includes surgery followed by radiation therapy. Surgery may be a lumpectomy or a mastectomy (e.g., total, simple or radical). Even if the doctor removes all of the cancer that can be seen at the time of surgery, the patient may be given radiation therapy, chemotherapy, or hormone therapy after surgery to try to kill any cancer cells that may be left. Radiation therapy is the use of x-rays or other types of radiation to kill cancer cells and shrink tumors. Radiation therapy may use external radiation (using a machine outside the body) or internal radiation. Chemotherapy is the use of drugs to kill cancer cells. Chemotherapy may be taken by mouth, or it may be put into the body by inserting a needle into a vein or muscle. Hormone therapy often focuses on estrogen and progesterone, which are hormones that affect the way some cancers grow. If tests show that the cancer cells have estrogen and progesterone receptors (molecules found in some cancer cells to which estrogen and progesterone will attach), hormone therapy is used to block the way these hormones help the cancer grow. Hormone therapy with tamoxifen is often given to patients with early stages of breast cancer and those with metastatic breast cancer. Other types of treatment being tested in clinical trials include sentinel lymph node biopsy followed by surgery and high-dose chemotherapy with bone marrow transplantation and peripheral blood stem cell transplantation. Any preventative/therapeutic treatment known in the art may be prescribed and/or administered, including, for example, surgery, chemotherapy and/or radiation treatment, and any of the treatments may be used in combination with one another to treat or prevent breast cancer (e.g., surgery followed by radiation therapy).

[0237] Also provided are methods of preventing or treating cancer comprising providing an individual in need of such treatment with a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* inhibitor that reduces or inhibits the overexpression of mutant *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* (e.g., a *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* polynucleotide with an allele that is associated with cancer). Included herein are methods of reducing or blocking the expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* comprising providing or administering to individuals in need of reducing or blocking the expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* a pharmaceutical or physiologically acceptable composition comprising a molecule capable of inhibiting expression of *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*, e.g., a siRNA molecule. Also included herein are methods of reducing or blocking the expression of secondary regulatory genes regulated by *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE* that play a role in oncogenesis which comprises introducing competitive

inhibitors that target *ICAM*, *MAPK10*, *KIAA0861*, *NUMA1* or *GALE*'s effect on these regulatory genes or that block the binding of positive factors necessary for the expression of these regulatory genes.

[0238] The examples set forth below are intended to illustrate but not limit the invention.

### Examples

[0239] In the following studies a group of subjects were selected according to specific parameters relating to breast cancer. Nucleic acid samples obtained from individuals in the study group were subjected to genetic analysis, which identified associations between breast cancer and certain polymorphic variants in *ICAM* region, *MAPK10*, *KIAA0861*, *NUMA1/FLJ20625/LOC220074* region, and *HT014/LOC148902/LYPLA2/GALE* region loci (herein referred to as "target genes", "target nucleotides", "target polypeptides" or simply "targets"). In addition, methods are described for combining information from multiple SNPs from the target genes found to be independently associated with breast cancer status in a case-control study. The resulting model permits a powerful, more informative quantitation of the combined value of the SNPs for predicting breast cancer susceptibility.

#### Example 1

##### Samples and Pooling Strategies

##### Sample Selection

[0240] Blood samples were collected from individuals diagnosed with breast cancer, which were referred to as case samples. Also, blood samples were collected from individuals not diagnosed with breast cancer as gender and age-matched controls. All of the samples were of German/German descent. A database was created that listed all phenotypic trait information gathered from individuals for each case and control sample. Genomic DNA was extracted from each of the blood samples for genetic analyses.

##### DNA Extraction from Blood Samples

[0241] Six to ten milliliters of whole blood was transferred to a 50 ml tube containing 27 ml of red cell lysis solution (RCL). The tube was inverted until the contents were mixed. Each tube was incubated for 10 minutes at room temperature and inverted once during the incubation. The tubes were then centrifuged for 20 minutes at 3000 x g and the supernatant was carefully poured off. 100-200 µl of residual liquid was left in the tube and was pipetted repeatedly to resuspend the pellet in the residual supernatant. White cell lysis solution (WCL) was added to the tube and pipetted repeatedly until completely mixed. While no incubation was normally required, the solution was incubated at 37°C or room temperature if cell clumps were visible after mixing until the solution was homogeneous. 2 ml of

protein precipitation was added to the cell lysate. The mixtures were vortexed vigorously at high speed for 20 sec to mix the protein precipitation solution uniformly with the cell lysate, and then centrifuged for 10 minutes at 3000 x g. The supernatant containing the DNA was then poured into a clean 15 ml tube, which contained 7 ml of 100% isopropanol. The samples were mixed by inverting the tubes gently until white threads of DNA were visible. Samples were centrifuged for 3 minutes at 2000 x g and the DNA was visible as a small white pellet. The supernatant was decanted and 5 ml of 70% ethanol was added to each tube. Each tube was inverted several times to wash the DNA pellet, and then centrifuged for 1 minute at 2000 x g. The ethanol was decanted and each tube was drained on clean absorbent paper. The DNA was dried in the tube by inversion for 10 minutes, and then 1000 µl of 1X TE was added. The size of each sample was estimated, and less TE buffer was added during the following DNA hydration step if the sample was smaller. The DNA was allowed to rehydrate overnight at room temperature, and DNA samples were stored at 2-8°C.

[0242] DNA was quantified by placing samples on a hematology mixer for at least 1 hour. DNA was serially diluted (typically 1:80, 1:160, 1:320, and 1:640 dilutions) so that it would be within the measurable range of standards. 125 µl of diluted DNA was transferred to a clear U-bottom microtitre plate, and 125 µl of 1X TE buffer was transferred into each well using a multichannel pipette. The DNA and 1X TE were mixed by repeated pipetting at least 15 times, and then the plates were sealed. 50 µl of diluted DNA was added to wells A5-H12 of a black flat bottom microtitre plate. Standards were inverted six times to mix them, and then 50 µl of 1X TE buffer was pipetted into well A1, 1000 ng/ml of standard was pipetted into well A2, 500 ng/ml of standard was pipetted into well A3, and 250 ng/ml of standard was pipetted into well A4. PicoGreen (Molecular Probes, Eugene, Oregon) was thawed and freshly diluted 1:200 according to the number of plates that were being measured. PicoGreen was vortexed and then 50µl was pipetted into all wells of the black plate with the diluted DNA. DNA and PicoGreen were mixed by pipetting repeatedly at least 10 times with the multichannel pipette. The plate was placed into a Fluoroskan Ascent Machine (microplate fluorometer produced by Labsystems) and the samples were allowed to incubate for 3 minutes before the machine was run using filter pairs 485 nm excitation and 538 nm emission wavelengths. Samples having measured DNA concentrations of greater than 450 ng/µl were re-measured for conformation. Samples having measured DNA concentrations of 20 ng/µl or less were re-measured for confirmation.

#### Pooling Strategies

[0243] Samples were placed into one of two groups based on disease status. The two groups were female case groups and female control groups. A select set of samples from each group were utilized to generate pools, and one pool was created for each group. Each individual sample in a pool was



represented by an equal amount of genomic DNA. For example, where 25 ng of genomic DNA was utilized in each PCR reaction and there were 200 individuals in each pool, each individual would provide 125 pg of genomic DNA. Inclusion or exclusion of samples for a pool was based upon the following criteria: the sample was derived from an individual characterized as Caucasian; the sample was derived from an individual of German paternal and maternal descent; the database included relevant phenotype information for the individual; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Phenotype information included pre- or post-menopausal, familial predisposition, country or origin of mother and father, diagnosis with breast cancer (date of primary diagnosis, age of individual as of primary diagnosis, grade or stage of development, occurrence of metastases, e.g., lymph node metastases, organ metastases), condition of body tissue (skin tissue, breast tissue, ovary tissue, peritoneum tissue and myometrium), method of treatment (surgery, chemotherapy, hormone therapy, radiation therapy). Samples that met these criteria were added to appropriate pools based on gender and disease status.

[0244] The selection process yielded the pools set forth in Table 1, which were used in the studies that follow:

**Table 1**

	<b>Female CASE</b>	<b>Female CONTROL</b>
<b>Pool size</b> (Number)	272	276
<b>Pool Criteria</b> (ex: case/control)	case	control
<b>Mean Age</b> (ex: years)	59.6	55.4

Example 2

Association of Polymorphic Variants with Breast cancer

[0245] A whole-genome screen was performed to identify particular SNPs associated with occurrence of breast cancer. As described in Example 1, two sets of samples were utilized, which included samples from female individuals having breast cancer (breast cancer cases) and samples from female individuals not having cancer (female controls). The initial screen of each pool was performed in an allelotyping study, in which certain samples in each group were pooled. By pooling DNA from each group, an allele frequency for each SNP in each group was calculated. These allele frequencies were then compared to one another. Particular SNPs were considered as being associated with breast cancer

when allele frequency differences calculated between case and control pools were statistically significant. SNP disease association results obtained from the allelotyping study were then validated by genotyping each associated SNP across all samples from each pool. The results of the genotyping were then analyzed, allele frequencies for each group were calculated from the individual genotyping results, and a p-value was calculated to determine whether the case and control groups had statistically significant differences in allele frequencies for a particular SNP. When the genotyping results agreed with the original allelotyping results, the SNP disease association was considered validated at the genetic level.

#### SNP Panel Used for Genetic Analyses

[0246] A whole-genome SNP screen began with an initial screen of approximately 25,000 SNPs over each set of disease and control samples using a pooling approach. The pools studied in the screen are described in Example 1. The SNPs analyzed in this study were part of a set of 25,488 SNPs confirmed as being statistically polymorphic as each is characterized as having a minor allele frequency of greater than 10%. The SNPs in the set reside in genes or in close proximity to genes, and many reside in gene exons. Specifically, SNPs in the set are located in exons, introns, and within 5,000 base-pairs upstream of a transcription start site of a gene. In addition, SNPs were selected according to the following criteria: they are located in ESTs; they are located in Locuslink or Ensemble genes; and they are located in Genomatix promoter predictions. SNPs in the set also were selected on the basis of even spacing across the genome, as depicted in Table 2.

[0247] A case-control study design using a whole genome association strategy involving approximately 28,000 single nucleotide polymorphisms (SNPs) was employed. Approximately 25,000 SNPs were evenly spaced in gene-based regions of the human genome with a median inter-marker distance of about 40,000 base pairs. Additionally, approximately 3,000 SNPs causing amino acid substitutions in genes described in the literature as candidates for various diseases were used. The case-control study samples were of female German origin (German paternal and maternal descent) 548 individuals were equally distributed in two groups (female controls and female cases). The whole genome association approach was first conducted on 2 DNA pools representing the 2 groups. Significant markers were confirmed by individual genotyping.

**Table 2**

General Statistics		Spacing Statistics	
Total # of SNPs	25,488	Median	37,058 bp
# of Exonic SNPs	>4,335 (17%)	Minimum*	1,000 bp
# SNPs with refSNP ID	20,776 (81%)	Maximum*	3,000,000 bp
Gene Coverage	>10,000	Mean	122,412 bp
Chromosome Coverage	All	Std Deviation	373,325 bp

		<i>*Excludes outliers</i>
--	--	---------------------------

### Allelotyping and Genotyping Results

[0248] The genetic studies summarized above and described in more detail below identified allelic variants associated with breast cancer. The allelic variants identified from the SNP panel described in Table 2 are summarized below in Table 3.

**Table 3**

SNP Reference	Chromosome Position	Position in Figure	Contig Identification	Contig Position	Sequence Identification	Locus	Sequence Position	Allelic Variability
1056538	10248147	44247	NT_011295		NM_000201	ICAM region		C/T
1541998	87342924	36424	NT_016354	11444849	NM_002753	MAPK10	intragenic	C/T
2001449	184330963	48563	NT_005962	18141399	NM_015078	KIAA0861	intragenic	G/C
673478	72021802	49002	NT_033927	1998133	NM_006185	NUMA1		T/C
					NM_017907	FLJ20625	downstream	
					NM_145309	LOC220074		
4237	10291777	87877	NT_004391	454476	NM_000403	GALE	downstream	A/G
			NT_004610		NM_020362	HT014		
			NO. INFO.		NO INFO.	LOC148902		
			NT_004610		NM_007260	LYPLA2		

[0249] Table 3 includes information pertaining to the incident polymorphic variant associated with breast cancer identified herein. Public information pertaining to the polymorphism and the genomic sequence that includes the polymorphism are indicated. The genomic sequences identified in Table 3 may be accessed at the http address [www.ncbi.nih.gov/entrez/query.fcgi](http://www.ncbi.nih.gov/entrez/query.fcgi), for example, by using the publicly available SNP reference number (e.g., rs1541998). The chromosome position refers to the position of the SNP within NCBI's Genome Build 33, which may be accessed at the following http address: [www.ncbi.nlm.nih.gov/mapview/map\\_search.cgi?chr=hum\\_chr.inf&query=](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=hum_chr.inf&query=). The "Contig Position" provided in Table 3 corresponds to a nucleotide position set forth in the contig sequence, and designates the polymorphic site corresponding to the SNP reference number. The sequence containing the polymorphisms also may be referenced by the "Sequence Identification" set forth in Table 3. The "Sequence Identification" corresponds to cDNA sequence that encodes associated target polypeptides (e.g., *NUMA1*) of the invention. The position of the SNP within the cDNA sequence is provided in the "Sequence Position" column of Table 3. Also, the allelic variation at the polymorphic site and the allelic variant identified as associated with breast cancer is specified in Table 3. All nucleotide sequences referenced and accessed by the parameters set forth in Table 3 are incorporated herein by reference. The positions for these SNPs are indicated in the tables below in Figures 1, 2, 3 and 4, and the incident SNP for the *GALE* region is at position 174 in Figure 5.

Assay for Verifying, Allelotyping, and Genotyping SNPs

[0250] A MassARRAY™ system (Sequenom, Inc.) was utilized to perform SNP genotyping in a high-throughput fashion. This genotyping platform was complemented by a homogeneous, single-tube assay method (hME™ or homogeneous MassEXTEND™ (Sequenom, Inc.)) in which two genotyping primers anneal to and amplify a genomic target surrounding a polymorphic site of interest. A third primer (the MassEXTEND™ primer), which is complementary to the amplified target up to but not including the polymorphism, was then enzymatically extended one or a few bases through the polymorphic site and then terminated.

[0251] For each polymorphism, SpectroDESIGNER™ software (Sequenom, Inc.) was used to generate a set of PCR primers and a MassEXTEND™ primer was used to genotype the polymorphism. Table 4 shows PCR primers and Table 5 shows extension primers used for analyzing polymorphisms. The initial PCR amplification reaction was performed in a 5 µl total volume containing 1X PCR buffer with 1.5 mM MgCl<sub>2</sub> (Qiagen), 200 µM each of dATP, dGTP, dCTP, dTTP (Gibco-BRL), 2.5 ng of genomic DNA, 0.1 units of HotStar DNA polymerase (Qiagen), and 200 nM each of forward and reverse PCR primers specific for the polymorphic region of interest.

**Table 4: PCR Primers**

Reference SNP ID	Forward PCR primer	Reverse PCR primer
1056538	GACAGCCACAGCTAGCGCAGA	TGTTTTCGCCCCCAGGGTGAC
1541998	CTGATTATTCTGATGGTAATG	GCCCATGTTAACATTTTCTTC
2001449	ATGTCAAGTGCACCCACATG	AGGAAGAACTGACGGAAGG
673478	TAATACAAAGGTGGCAGCAG	TTGACAAGGATAAGGACAAG
4237	GCACATGGCCACATTAAGTGG	TGGCTGTGGAAATTGGGTCTTG

[0252] Samples were incubated at 95°C for 15 minutes, followed by 45 cycles of 95°C for 20 seconds, 56°C for 30 seconds, and 72°C for 1 minute, finishing with a 3 minute final extension at 72°C. Following amplification, shrimp alkaline phosphatase (SAP) (0.3 units in a 2 µl volume) (Amersham Pharmacia) was added to each reaction (total reaction volume was 7 µl) to remove any residual dNTPs that were not consumed in the PCR step. Samples were incubated for 20 minutes at 37°C, followed by 5 minutes at 85°C to denature the SAP.

[0253] Once the SAP reaction was complete, a primer extension reaction was initiated by adding a polymorphism-specific MassEXTEND™ primer cocktail to each sample. Each MassEXTEND™ cocktail included a specific combination of dideoxynucleotides (ddNTPs) and deoxynucleotides (dNTPs)

used to distinguish polymorphic alleles from one another. In Table 5, ddNTPs are shown and the fourth nucleotide not shown is the dNTP.

**Table 5: Extend Primers**

Reference SNP ID	Extend Probe	Term Mix
1056538	CCCAGGGTGACGTTGCAGA	ACG
1541998	ATTATTCTGATGGTAATGATCCAG	ACG
2001449	CACATGCCTGCTCGCCCCC	ACT
673478	AAGGGGAGGTCGACTGGG	ACT
4237	GGCATCTGGCAGTCATGG	ACT

**[0254]** The MassEXTEND™ reaction was performed in a total volume of 9 µl, with the addition of 1X ThermoSequenase buffer, 0.576 units of ThermoSequenase (Amersham Pharmacia), 600 nM MassEXTEND™ primer, 2 mM of ddATP and/or ddCTP and/or ddGTP and/or ddTTP, and 2 mM of dATP or dCTP or dGTP or dTTP. The deoxy nucleotide (dNTP) used in the assay normally was complementary to the nucleotide at the polymorphic site in the amplicon. Samples were incubated at 94°C for 2 minutes, followed by 55 cycles of 5 seconds at 94°C, 5 seconds at 52°C, and 5 seconds at 72°C.

**[0255]** Following incubation, samples were desalted by adding 16 µl of water (total reaction volume was 25 µl), 3 mg of SpectroCLEAN™ sample cleaning beads (Sequenom, Inc.) and allowed to incubate for 3 minutes with rotation. Samples were then robotically dispensed using a piezoelectric dispensing device (SpectroJET™ (Sequenom, Inc.)) onto either 96-spot or 384-spot silicon chips containing a matrix that crystallized each sample (SpectroCHIP® (Sequenom, Inc.)). Subsequently, MALDI-TOF mass spectrometry (Biflex and Autoflex MALDI-TOF mass spectrometers (Bruker Daltonics) can be used) and SpectroTYPER RT™ software (Sequenom, Inc.) were used to analyze and interpret the SNP genotype for each sample.

#### Genetic Analysis

**[0256]** Variations identified in the target genes are provided in their respective genomic sequences (see Figures 1-5) Minor allelic frequencies for these polymorphisms was verified as being 10% or greater by determining the allelic frequencies using the extension assay described above in a group of samples isolated from 92 individuals originating from the state of Utah in the United States, Venezuela and France (Coriell cell repositories).

[0257] Genotyping results are shown for female pools in Table 6A and 6B. Table 6A shows the original genotyping results and Table 6B shows the genotyped results re-analyzed to remove duplicate individuals from the cases and controls (*i.e.*, individuals who were erroneously included more than once as either cases or controls). Therefore, Table 6B represents a more accurate measure of the allele frequencies for this particular SNP. In the subsequent tables, “AF” refers to allelic frequency; and “F case” and “F control” refer to female case and female control groups, respectively.

**Table 6A**

Reference SNP ID	AF F case	AF F control	p-value	Odds Ratio	Breast Cancer Assoc. Allele
1056538	C = 0.651 T = 0.349	C = 0.564 T = 0.436	<b>0.0038</b>	0.69	C
1541998	T = 0.780 C = 0.220	T = 0.839 C = 0.161	<b>0.0153</b>	0.69	C
2001449	G = 0.703 C = 0.297	G = 0.780 C = 0.220	<b>0.0040</b>	1.49	C
673478	T = 0.919 C = 0.081	T = 0.953 C = 0.047	<b>0.0238</b>	1.74	C
4237	A = 0.590 G = 0.410	A = 0.530 G = 0.470	<b>0.0431</b>	0.78	A

**Table 6B**

Reference SNP ID	AF F case	AF F control	p-value	Odds Ratio	Breast Cancer Assoc. Allele
1056538	C = 0.658 T = 0.342	C = 0.556 T = 0.444	<b>0.0012</b>	0.65	C
1541998	T = 0.771 C = 0.229	T = 0.839 C = 0.161	<b>0.0070</b>	0.65	C
2001449	G = 0.693 C = 0.307	G = 0.782 C = 0.218	<b>0.0012</b>	1.59	C
673478	T = 0.916 C = 0.084	T = 0.953 C = 0.047	<b>0.0171</b>	1.85	C
4237	A = 0.584 G = 0.416	A = 0.527 G = 0.473	0.0704	0.79	A

[0258] The single marker alleles set forth in Table 3 were considered validated, since the genotyping data for the females, males or both pools were significantly associated with breast cancer, and because the genotyping results agreed with the original allelotyping results. Particularly significant associations with breast cancer are indicated by a calculated p-value of less than 0.05 for genotype results, which are set forth in bold text.

[0259] Odds ratio results are shown in Tables 6A and 6B. An odds ratio is an unbiased estimate of relative risk which can be obtained from most case-control studies. Relative risk (RR) is an estimate of the likelihood of disease in the exposed group (susceptibility allele or genotype carriers) compared to the unexposed group (not carriers). It can be calculated by the following equation:

$$RR = IA/Ia$$

$IA$  is the incidence of disease in the A carriers and  $Ia$  is the incidence of disease in the non-carriers.

**$RR > 1$  indicates the A allele increases disease susceptibility.**

$RR < 1$  indicates the a allele increases disease susceptibility.

For example,  $RR = 1.5$  indicates that carriers of the A allele have 1.5 times the risk of disease than non-carriers, *i.e.*, 50% more likely to get the disease.

[0260] Case-control studies do not allow the direct estimation of  $IA$  and  $Ia$ , therefore relative risk cannot be directly estimated. However, the odds ratio (OR) can be calculated using the following equation:

$$OR = (nDA/ndA)/(nDa/ndA) = pDA(1 - pdA)/pdA(1 - pDA), \text{ or}$$

$$OR = ((\text{case } f) / (1 - \text{case } f)) / ((\text{control } f) / (1 - \text{control } f)), \text{ where } f = \text{susceptibility allele frequency.}$$

[0261] An odds ratio can be interpreted in the same way a relative risk is interpreted and can be directly estimated using the data from case-control studies, *i.e.*, case and control allele frequencies. The higher the odds ratio value, the larger the effect that particular allele has on the development of breast cancer. Possessing an allele associated with a relatively high odds ratio translates to having a higher risk of developing or having breast cancer.

### Example 3

#### Samples and Pooling Strategies for the Replication Samples

[0262] The SNPs of Table 3 were genotyped again in a collection of replication samples to further validate its association with breast cancer. Like the original study population described in Examples 1 and 2, the replication samples consisted of females diagnosed with breast cancer (cases) and females without cancer (controls). The case and control samples were selected and genotyped as described below.

#### Pooling Strategies

[0263] Samples were placed into one of two groups based on disease status. The two groups were female case groups and female control groups. A select set of samples from each group were utilized to generate pools, and one pool was created for each group. Each individual sample in a pool was represented by an equal amount of genomic DNA. For example, where 25 ng of genomic DNA was

utilized in each PCR reaction and there were 190 individuals in each pool (i.e., 190 cases and 190 controls), each individual would provide 125 pg of genomic DNA. Inclusion or exclusion of samples for a pool was based upon the following criteria: the sample was derived from a female individual characterized as Caucasian from Australia; case samples were derived from individuals diagnosed with breast cancer; control samples were derived from individuals free of cancer and no family history of breast cancer; and sufficient genomic DNA was extracted from each blood sample for all allelotyping and genotyping reactions performed during the study. Samples in the pools also were age-matched. Samples that met these criteria were added to appropriate pools based on gender and disease status.

[0264] The selection process yielded the pools set forth in Table 7, which were used in the studies that follow:

**Table 7**

	<b>Female CASE</b>	<b>Female CONTROL</b>
<b>Pool size</b> (Number)	190	190
<b>Pool Criteria</b> (ex: case/control)	Case	control
<b>Mean Age</b> (ex: years)	64.5	**

\*\*Each case was matched by a control within 5 years of age of the case.

[0265] The replication genotyping results are shown in Table 8. The odds ratio was calculated as described in Example 2.

**Table 8**

<b>Reference SNP ID</b>	<b>AF F case</b>	<b>AF F control</b>	<b>p-value</b>	<b>Odds Ratio</b>
1056538	C = 0.650 T = 0.350	C = 0.584 T = 0.416	<b>0.0624</b>	0.75
1541998	T = 0.820 C = 0.180	T = 0.864 C = 0.136	0.1010	0.72
2001449	G = 0.685 C = 0.315	G = 0.777 C = 0.223	<b>0.005</b>	1.59
673478	T = 0.927 C = 0.073	T = 0.957 C = 0.043	<b>0.077</b>	1.76
4237	A = 0.632 G = 0.368	A = 0.577 G = 0.423	0.1260	1.26

[0266] The absence of a statistically significant association in the replication cohort should not be interpreted as minimizing the value of the original finding. There are many reasons why a biologically derived association identified in a sample from one population would not replicate in a sample from another population. The most important reason is differences in population history. Due to bottlenecks



and founder effects, there may be common disease predisposing alleles present in one population that are relatively rare in another, leading to a lack of association in the candidate region. Also, because common diseases such as breast cancer are the result of susceptibilities in many genes and many environmental risk factors, differences in population-specific genetic and environmental backgrounds could mask the effects of a biologically relevant allele. For these and other reasons, statistically strong results in the original, discovery sample that did not replicate in the replication sample may be further evaluated in additional replication cohorts and experimental systems.

#### Example 4

##### ICAM Region Proximal SNPs

[0267] It has been discovered that a polymorphic variation (rs1056538) in a region that encodes ICAM1, ICAM2 and ICAM5 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs1056538) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately seventy-five allelic variants located within the ICAM region were identified and allelotyped. The polymorphic variants are set forth in Table 9. The chromosome position provided in column four of Table 9 is based on Genome “Build 33” of NCBI’s GenBank.

**Table 9**

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
2884487	19	139	10204039	T/C
1059840	19	11799	10215699	A/T
11115	19	11851	10215751	T/C
1059849	19	11963	10215863	G/A
3093035	19	24282	10228182	A/G
ICAM_SNPA	19	26849	10230749	A/T
281428	19	29633	10233533	C/T
281431	19	31254	10235154	T/C
ICAM_SNPB	19	31967	10235867	G/C
2358581	19	32920	10236820	G/T
281434	19	33929	10237829	A/G
ICAM_SNPC	19	35599	10239499	G/C
1799969	19	36101	10240001	G/A
3093033	19	36340	10240240	G/A
ICAM_SNPD	19	36405	10240305	A/G
ICAM_SNPE	19	36517	10240417	T/C
ICAM_SNPF	19	36777	10240677	A/G
5498	19	36992	10240892	G/A
ICAM_SNPG	19	37645	10241545	T/C
1057981	19	37868	10241768	G/A
281436	19	38440	10242340	A/G

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
923366	19	38532	10242432	T/C
281437	19	38547	10242447	C/T
ICAM_SNP	19	38712	10242612	T/C
281438	19	40684	10244584	T/G
3093029	19	40860	10244760	C/G
2569693	19	41213	10245113	C/T
281439	19	41419	10245319	G/C
281440	19	41613	10245513	G/A
ICAM_SNP	19	42407	10246307	C/G
1333881	19	43440	10247340	T/C
1056538	19	44247	10248147	T/C
2228615	19	44677	10248577	A/G
2569702	19	45256	10249156	T/C
2569703	19	45536	10249436	C/G
ICAM_SNP	19	46153	10250053	C/T
2569707	19	47546	10251446	C/G
2916060	19	47697	10251597	A/C
885743	19	47944	10251844	A/T
ICAM_SNP	19	48530	10252430	C/G
892188	19	51102	10255002	T/C
2291473	19	57090	10260990	T/C
281416	19	60093	10263993	A/G
281417	19	60439	10264339	T/C
281418	19	62694	10266594	G/C
430092	19	66260	10270160	C/T
368835	19	67295	10271195	A/G
2358583	19	67304	10271204	T/G
ICAM_SNP	19	67731	10271631	G/T
1045384	19	68555	10272455	C/A
281427	19	70429	10274329	C/T
3745264	19	70875	10274775	T/G
281426	19	72360	10276260	G/A
281424	19	74228	10278128	C/T
281423	19	76802	10280702	C/T
281422	19	77664	10281564	T/C
281421	19	78803	10282703	A/G
281420	19	79263	10283163	A/G
3745263	19	80810	10284710	A/G
3745261	19	81020	10284920	T/C
3181049	19	82426	10286326	T/C
281412	19	82783	10286683	T/C
2230399	19	85912	10289812	C/G
2278442	19	86135	10290035	G/A
2304237	19	87877	10291777	T/C
281413	19	88043	10291943	G/A
1058154	19	88206	10292106	A/C
3176769	19	88343	10292243	T/C

dbSNP rs#	Chromosome	Position in Figure 1	Chromosome Position	Allele Variants
2304240	19	90701	10294601	G/A
3176768	19	90974	10294874	A/G
3176767	19	91060	10294960	C/A
3176766	19	91087	10294987	C/T
ICAM_SNP	19	91594	10295494	G/A
281415	19	92302	10296202	T/G
3176764	19	92384	10296284	A/G

#### Assay for Verifying and Allelotyping SNPs

[0268] The methods used to verify and allelotype the proximal SNPs of Table 9 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 10 and Table 11, respectively.

**Table 10**

dbSNP rs#	Forward PCR primer	Reverse PCR primer
5498	ACGTTGGATGCTCACAGAGCACATTCACGG	ACGTTGGATGAGATCTTGAGGGCACCTACC
11115	ACGTTGGATGAGGTGACACCTTCCTCGAAG	ACGTTGGATGTGTGAAGCACCTCTTCTGAG
11115	ACGTTGGATGGTCCAGGTGACACCTTCCTC	ACGTTGGATGAAGCACCTCTTCTGAGCCAG
56901	ACGTTGGATGGTCCAGGTGACACCTTCCTC	ACGTTGGATGAAGCACCTCTTCTGAGCCAG
240914	ACGTTGGATGTTCAACAAGCGAGTGACAGC	ACGTTGGATGGTGCAGAGATGGGCTTTCTC
254615	ACGTTGGATGTGTAGATGGTCACGTTCTCC	ACGTTGGATGATCTGAGTCCTGATGTCACC
254615	ACGTTGGATGTTGCAGCTTTAAGCTAAGGC	ACGTTGGATGAGCCCAGGAGACTTAATTAC
272539	ACGTTGGATGTACAGACCCCTCTACCCCTTC	ACGTTGGATGAGGTGACACCTTCCTCGAAG
281412	ACGTTGGATGTGACCTCAGGTGATTCACCC	ACGTTGGATGGGTATACCTTTAGCTGGCTG
281413	ACGTTGGATGTCAAAGCTCACAGTTCTCGG	ACGTTGGATGACTTAGCGGGTCTGCAAAC
281414	ACGTTGGATGAAGGCACCTTCCTCTGTCAG	ACGTTGGATGTGGGCCACAACACGGATGGTA
281415	ACGTTGGATGGCACAAGAGCTAAGGTAGG	ACGTTGGATGGAATCCTGGATAGACAGTGG
281416	ACGTTGGATGTAACGTAGAGCACAGGTGAG	ACGTTGGATGCAACGCAAACACCAGTGTGG
281417	ACGTTGGATGAAGAGACAGTGGAGAGGCTG	ACGTTGGATGAGAGCCATCGGGTCCCAGCAA
281418	ACGTTGGATGTGCGCTCAGTCAGCTTCCTC	ACGTTGGATGAGTGTAGCCGAGGGCAAGC
281420	ACGTTGGATGCCAGGACTGTCTCTCTGTTT	ACGTTGGATGATGACACTACAGCCTGAGCA
281421	ACGTTGGATGAGTGTGCTTTGTACCCAG	ACGTTGGATGAGGAGAATCGCTTGACCTG
281422	ACGTTGGATGAGAAATCCTCCTACCTTGCC	ACGTTGGATGGCCCGGCTCTACATAAAAT
281423	ACGTTGGATGAACCTCAAGCTGCTTCACTG	ACGTTGGATGGAGGAGCCACCTTTAATGT
281424	ACGTTGGATGACCTGTGTTTCTAGGTGTGC	ACGTTGGATGCATGCCTGGGAAAAAACTCC
281426	ACGTTGGATGATCCTCACACCTCAGTCTCC	ACGTTGGATGAATGAGACTCCGTCTCTACC
281427	ACGTTGGATGGACAATTGTAGTACCCAGCC	ACGTTGGATGAGGAGAATCGCTTGAACCTG
281428	ACGTTGGATGAGTAGCTGGAATTACAGGCG	ACGTTGGATGGCCAAACATGATGAAATCCCG
281431	ACGTTGGATGACTGGGATTACAGGTGTGAG	ACGTTGGATGGGAGAAATCTTGATGGAGGC
281432	ACGTTGGATGAGCTGGGACTTTCCTTCTTG	ACGTTGGATGCAGTAAATCCAGCCTTCAGC
281434	ACGTTGGATGCCACGCCTGGCTAATTTTTG	ACGTTGGATGGGTGAGGAGTTCAAGACCAG
281436	ACGTTGGATGCATGGTTCACTGCAGTCTTG	ACGTTGGATGTGTGGTGTGTGAGCCTATG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
281437	ACGTTGGATGATAGGCTCACAAACACCACAC	ACGTTGGATGAACACAAAGGAAGTCTGGGC
281437	ACGTTGGATGATAGGCTCACAAACACCACAC	ACGTTGGATGAACACAAAGGAAGTCTGGGC
281438	ACGTTGGATGACCTGAGGTTTCCTCACTCAG	ACGTTGGATGAGAGGTTTCTGTGACACCCG
281439	ACGTTGGATGGCGGAGCCATACCTCTAAGC	ACGTTGGATGTCGCTGGCACTTTCGTCCC
281440	ACGTTGGATGCTGGCTGAGATGCCATGATA	ACGTTGGATGATGGTGGGAGGAGCTAAATG
281440	ACGTTGGATGGCCATGATAATAAGCTGGAC	ACGTTGGATGTCTTAGTCCCCAAATGTATC
368835	ACGTTGGATGGGTGGGAAAAAGACGTGAAG	ACGTTGGATGAGAGGGAATTAAGGAGGTCC
378395	ACGTTGGATGAATTCCGTGGGATGAGGAAT	ACGTTGGATGACCGTGTTTTCCAGGCTCGCG
378395	ACGTTGGATGACTTGGCCCCCTGCACTCACA	ACGTTGGATGACCGTGTTTTCCAGGCTCGCG
430092	ACGTTGGATGGTTGGGATTACAGGCATGAG	ACGTTGGATGATCTGTTGCCTGTCAAGATG
473241	ACGTTGGATGGCCATGATAATAAGCTGGAC	ACGTTGGATGAAATGTATCCCCGCCCTAAG
547878	ACGTTGGATGTACTCAGGAGGCTGAGGTG	ACGTTGGATGCATGGTTCACTGCAGTCTTG
827786	ACGTTGGATGGCGGAGCCATACCTCTAAGC	ACGTTGGATGTCGCTGGCACTTTCGTCCC
827787	ACGTTGGATGCTGGCTGAGATGCCATGATA	ACGTTGGATGATGGTGGGAGGAGCTAAATG
885743	ACGTTGGATGTGAGAGAAGGCGATCTTGAC	ACGTTGGATGCCAATTCACAATCCACTGTG
885743	ACGTTGGATGTGAGAGAAGGCGATCTTGAC	ACGTTGGATGCCAATTCACAATCCACTGTG
892188	ACGTTGGATGGTTTGTTTTAGAGACAGGG	ACGTTGGATGGTCAAAGCCACTTCCAGCTA
901886	ACGTTGGATGCGATCTGGTCTGCTCTGCAAG	ACGTTGGATGGCCCCACCTTCTGTTCCAAG
923366	ACGTTGGATGTCTGGGCAATGTTGCAAGAC	ACGTTGGATGATAGGCTCACAAACACCACAC
923366	ACGTTGGATGTCTGGGCAATGTTGCAAGAC	ACGTTGGATGATAGGCTCACAAACACCACAC
1045384	ACGTTGGATGGTGCAGAGATGGGCTTCTC	ACGTTGGATGAGATGGGCACAATGTCCGAC
1056538	ACGTTGGATGACTGCCACAGCCACAGCTAG	ACGTTGGATGTTTTCGCCCCCAGGGTGA
1057981	ACGTTGGATGGTACAACCTGTACCTGGTGAC	ACGTTGGATGAATGAACATAGGTCTCTGGC
1058154	ACGTTGGATGTCCCTTCCATCCTCATTTTT	ACGTTGGATGTGCAAGGCGCTAAACAAAAC
1059840	ACGTTGGATGTGCGCCTGGCTCAGAAGAGG	ACGTTGGATGACCCCTACCCACGCTACCCA
1059849	ACGTTGGATGGGAATGGATGCAGAAGCCCG	ACGTTGGATGAAGCTGAGGCCACAGGGAG
1059849	ACGTTGGATGAATGGATGCAGAAGCCCGTC	ACGTTGGATGATTCCACGGAGGAAGCTGAG
1333881	ACGTTGGATGATCAGCTCTACGCGATCTGG	ACGTTGGATGTTACGGCCCCACCTTCTGTTT
1799969	ACGTTGGATGTCAACCTCTGGTCCCCCAGTG	ACGTTGGATGAGGGGACCGTGGTCTGTTT
1799969	ACGTTGGATGTTGCCATAGGTGACTGTGGG	ACGTTGGATGTCCTAGAGGTGGACACGCAG
2075741	ACGTTGGATGAAGATGCCAGTCCGTGGACC	ACGTTGGATGCTGGAGACCCAGTGCTCTC
2228615	ACGTTGGATGGGGCAGATGGTGACAGTAAC	ACGTTGGATGTGGAACCTCCCTCCAGTGTGA
2228615	ACGTTGGATGGGGCAGATGGTGACAGTAAC	ACGTTGGATGTGGAACCTCCCTCCAGTGTGA
2230399	ACGTTGGATGAGCGGCAGTTACCATGTTAG	ACGTTGGATGTTCTTCCCCCATTGCTTCTG
2230399	ACGTTGGATGAGCGGCAGTTACCATGTTAG	ACGTTGGATGTTCTTCCCCCATTGCTTCTG
2278442	ACGTTGGATGGGTGATGGACATTGAGGGTG	ACGTTGGATGTCCCTTCTGTCTCCAACCC
2278442	ACGTTGGATGTCGTGGTGTGACATTGAG	ACGTTGGATGAAGTCAATATGCGTCCCTTC
2291473	ACGTTGGATGAAGAGGCTATGTGGCAGATG	ACGTTGGATGAGGGTGAAGCTGGGTTTAAC
2304237	ACGTTGGATGTGGGCCAGAACTTACCCTG	ACGTTGGATGAAGCAGCACCACCGTGAGG
2304240	ACGTTGGATGAATCTCAGCAACGTGACTGG	ACGTTGGATGACACGGTGTGTTAGAGGAG
2304240	ACGTTGGATGAATCTCAGCAACGTGACTGG	ACGTTGGATGACACGGTGTGTTAGAGGAG
2358581	ACGTTGGATGTAAGGCAGGAGGATGGAGTG	ACGTTGGATGGACAGAGTCTCACTCTGTCTG
2358583	ACGTTGGATGAAGACGTGAAGAGACACACC	ACGTTGGATGAGAGGGAATTAAGGAGGTCC
2569693	ACGTTGGATGCTTGTTCTCGCGTGGATGTC	ACGTTGGATGTACTCAGCGTGTGTGAGCTC
2569702	ACGTTGGATGACCCTCCAGACCTTGAACCA	ACGTTGGATGACGTAACGCTAACGGTGGAG
2569702	ACGTTGGATGATACCCTACTCCTACTCTTC	ACGTTGGATGTCAAGGACGTAACGCTAACG
2569703	ACGTTGGATGTCAGGAAGCTCCCAGACAGA	ACGTTGGATGATAACCTTGGACGCCGATC

dbSNP rs#	Forward PCR primer	Reverse PCR primer
2569703	ACGTTGGATGTTAGACGAAAAAGGCGCCAC	ACGTTGGATGTTGTCCCTGCATAACCCTTG
2569707	ACGTTGGATGTGAGCGTGGCAGGCGCCATG	ACGTTGGATGGCGTGGCGCCCGTGCGCGT
2884487	ACGTTGGATGTGTGGCAAATGATGGAACAG	ACGTTGGATGCCAGAAGTTTGAGATCTGCC
2916060	ACGTTGGATGGGCGAGGTATCTGAGAGGG	ACGTTGGATGTACTCTGTCCCACCTCCGTC
3093029	ACGTTGGATGGGCAGCTCTGATTGGATGTT	ACGTTGGATGCTCCACAGTTGTTTGGCCTC
3093030	ACGTTGGATGAGAGACCCAGAAGGTCATAG	ACGTTGGATGCCTCCCCCAAGAAAACATTG
3093032	ACGTTGGATGGGCCACTTCTTCTGTAAGTC	ACGTTGGATGCATGAGGACATACAACCTGGG
3093033	ACGTTGGATGAAAGCCTGGAATAGGCACAC	ACGTTGGATGTGCAGACAGTGACCATCTAC
3093035	ACGTTGGATGGGAGACATAGCGAGATTCTG	ACGTTGGATGTAGAAAGCAGTGCGATCTGG
3176764	ACGTTGGATGAAATCGTTTGAACCCGGGAG	ACGTTGGATGGTTTTGAGACAGAGTCTCAC
3176766	ACGTTGGATGTTTCGGGCTGCAATGGTCCC	ACGTTGGATGTAACACCTCTCTCCTTGTC
3176767	ACGTTGGATGCGGTCTCTGATGGATTCTAC	ACGTTGGATGAACAGGCCCCACCATTTAAC
3176768	ACGTTGGATGGAGAGGTGTTAAATGGTGGG	ACGTTGGATGGGAACATGAAGAAGTCCTGG
3176769	ACGTTGGATGTTCTGTATTATGGCCAGACG	ACGTTGGATGGTCTGAACCTGATTGGAGAG
3181049	ACGTTGGATGATCTTCAGGGATGGTCACTC	ACGTTGGATGGACAAATACAAAGGGACAGG
3745261	ACGTTGGATGACACACAGCAGGGCATCCGT	ACGTTGGATGCGCAATCAATGCTTTCCACC
3745263	ACGTTGGATGTACATGAAGAAGGACTCGGC	ACGTTGGATGATCCGTCCAGTGACAGTAGA
3745264	ACGTTGGATGCAAAGTGCTAGGATCACAGG	ACGTTGGATGACTGCCCCATAGAGTGGCAA
FCH-0994	ACGTTGGATGTTTTCGCCCCCAGGGTGAC	ACGTTGGATGACAGCCACAGCTAGCGCAGA

Table 11

dbSNP rs#	Extend Primer	Term Mix
5498	CAGAGCACATTCACGGTCACCT	CGT
11115	AAGGGTGGGCGTGGGCCT	ACT
11115	AAGGGTGGGCGTGGGCCT	ACT
56901	AAGGGTGGGCGTGGGCCT	ACT
240914	ACAATGTCCGACTCCCACA	ACT
254615	CCAGGGTGACGTTGCAGA	ACG
254615	TAAGGCAAAGTTCAGCTACTTA	CGT
272539	ACCCCGTACCACTGTTGA	CGT
281412	GCTGGGATTATAAGCGTG	ACT
281413	GCTCACAGTTCTCGGCAGGAC	ACG
281414	CCTTCCTCTGTCAGAATGGC	ACG
281415	GGTGATTTGGGGACAGCTGA	ACT
281416	GGTCCACACCGACGCCAG	ACT
281417	CCCCTGCCAGGACACCCC	ACT
281418	TCAGCTTCCTCCCTCCCC	ACT
281420	ACTGTCTCTGTGTTTTGAGAT	ACT
281421	GCTTTGTCACCCAGGCTGGA	ACT
281422	CTGGGGAACACAGGAATGC	ACT
281423	GCCACCCTCCATTACAGC	ACG
281424	TAGGTGTGCGTGTGTGTGTG	ACG
281426	GAGCTGGGACCACAGGCA	ACG

dbSNP rs#	Extend Primer	Term Mix
281427	CTTTGTATACAATCTTCCCTC	ACG
281428	GCGCCCAGCACCACGCC	ACG
281431	ACAGGTGTGAGCCACTGC	ACT
281432	GGGAGTCATGGAGGGTTT	ACT
281434	TAGAGACGGGGTTTCACTAT	ACT
281436	ACTGCAGTCTTGACCTTTTG	ACT
281437	TTTTTTTTCCAGAGACGGGGTCT	ACG
281437	TTTTTCCAGAGACGGGGTCT	ACG
281438	CGAAGCCCCAGACTCTGTGTA	ACT
281439	ACCCCTCCGGGTCAGCTCC	ACT
281440	TAATAAGCTGGACTCCGAGC	ACG
281440	TAATAAGCTGGACTCCGAGC	ACG
368835	AGACGTGAAGAGACACACCT	ACT
378395	GCCCGCGTCCTCCTCTCC	ACT
378395	GCCCGCGTCCTCCTCTCC	ACT
430092	ATTACAGGCATGAGCCACTG	ACG
473241	ATAATAAGCTGGACTCCGAGC	ACG
547878	GTGGGAGGATCACTTGAGC	ACG
827786	ACCCCTCCGGGTCAGCTCC	ACT
827787	TAATAAGCTGGACTCCGAGC	ACG
885743	GACCCCTCTCTCCCTCCA	CGT
885743	GACCCCTCTCTCCCTCCA	CGT
892188	TGGGCTGGAGCACAATGAC	ACT
901886	GAGTCCGCAGCTCTTTGAAC	ACT
923366	TTGCAAGACCCCGTCTCTG	ACT
923366	TTGCAAGACCCCGTCTCTG	ACT
1045384	CCAGTCCCCTGCTGTCTGT	CGT
1056538	GAGGGTGCCAGGCAGCTG	ACT
1057981	TACCTGGTGACCTTGAATGTGAT	ACG
1058154	CTTCCATCCTCATTTTTTTTATT	ACT
1059840	GCTCAGAAGAGGTGCTTCAC	CGT
1059849	CAGAAGCCCGTCTGGGCT	ACG
1059849	CAGAAGCCCGTCTGGGCT	ACG
1333881	AGAGTCCGCAGCTCTTTGAAC	ACT
1799969	CCGAGACTGGGAACAGCC	ACG
1799969	CCGAGACTGGGAACAGCC	ACG
2075741	GGACCATGGTGCACAGCA	ACT
2228615	AGTAACCTGCGCAGCTGGG	ACT
2228615	GTAACCTGCGCAGCTGGG	ACT
2230399	GTTACCATGTTAGGGAGGAGA	ACT
2230399	ACCATGTTAGGGAGGAGA	ACT
2278442	GGACATTGAGGGTGAGCTAA	ACG
2278442	ACATTGAGGGTGAGCTAA	ACG
2291473	GGAGTGTCCCTGGACCCC	ACT

dbSNP rs#	Extend Primer	Term Mix
2304237	TGCGCTGCCAAGTGGAGG	ACT
2304240	GCTCAGTGTACTGCAATGGCTC	ACG
2304240	AGTGTACTGCAATGGCTC	ACG
2358581	CTTGCACTGAGCCCAGATCG	CGT
2358583	AAGAGACACACCTAATTTGTGG	ACT
2569693	CGCGTGGATGTCAGGGCC	ACG
2569702	CAGACCTTGAACCAGATAGAA	ACT
2569702	ACCTTGAACCAGATAGAA	ACT
2569703	CTCCCAGACAGAGTGCATG	ACT
2569703	TCCCAGACAGAGTGCATG	ACT
2569707	GGCGAGTACGAGTGCACA	ACT
2884487	AGAGACAGGGTCTCGCC	ACT
2916060	CTCCCTCTCGGTCCCGG	ACT
3093029	AGTTTCCTATCCCAGCC	ACT
3093030	CCAGAACCTCAGGGTATG	
3093032	CTTCTGTAAGTCTGTGGG	
3093033	GGGTTGAGTGCACACC	ACG
3093035	TTCTGTCTCAAAAAACAAAGC	ACT
3176764	CCCGCCACTGCACTCCA	ACT
3176766	TCCTTCTGAGTTCTCCC	ACG
3176767	TGGATTCTACCTTTCCC	CGT
3176768	TGTTGATGCGTGGGTTGGGG	ACT
3176769	CGGGGTGGGTGGATCAA	ACT
3181049	ACTCCCTGCCCTGGCCC	ACT
3745261	GCAGCTGCACCGACAGTTC	ACT
3745263	TCGGCTGCCCCGTGCCAAGTC	ACT
3745264	ATACCATGCCAGGCATT	ACT
FCH-0994	CCCAGGGTGACGTTGCAGA	ACG

#### Genetic Analysis of Allelotyping Results

[0269] Allelotyping results are shown for cases and controls in Table 12. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where “AF” is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs2884487 has the following case and control allele frequencies: case A1 (T) = 0.788; case A2 (C) = 0.212; control A1 (T) = 0.758; and control A2 (C) = 0.242, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

Table 12

dbSNP rs#	Position in Figure 1	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2884487	139	10204039	T/C	0.212	0.242	0.2425
1059840	11799	10215699	A/T	0.809	0.805	0.8545
11115	11851	10215751	T/C	0.434	0.379	0.0644
1059849	11963	10215863	G/A	0.243	0.194	0.0468
3093035	24282	10228182	A/G	0.889	0.914	0.1592
ICAM SNPA	26849	10230749	A/T	Not Allelotyped		
281428	29633	10233533	C/T	0.180	0.174	0.7908
281431	31254	10235154	T/C	0.107	0.109	0.8964
ICAM SNPB	31967	10235867	G/C	0.375	0.382	0.8113
2358581	32920	10236820	G/T	0.097	0.074	0.1800
281434	33929	10237829	A/G	0.818	0.831	0.5765
ICAM SNPC	35599	10239499	G/C	Not Allelotyped		
1799969	36101	10240001	G/A	0.117	0.151	0.1036
3093033	36340	10240240	G/A	0.004	0.023	0.0051
ICAM SNPD	36405	10240305	A/G	Not Allelotyped		
ICAM SNPE	36517	10240417	T/C	Not Allelotyped		
ICAM SNPF	36777	10240677	A/G	Not Allelotyped		
5498	36992	10240892	G/A	0.554	0.487	0.0257
ICAM SNPG	37645	10241545	T/C	0.684	0.732	0.0788
1057981	37868	10241768	G/A	0.978	0.994	0.0289
281436	38440	10242340	A/G	0.504	0.554	0.0977
923366	38532	10242432	T/C	0.597	0.553	0.1471
281437	38547	10242447	C/T	0.195	0.151	0.0521
ICAM SNPH	38712	10242612	T/C	0.448	0.398	0.0970
281438	40684	10244584	T/G	0.235	0.200	0.1589
3093029	40860	10244760	C/G	0.089	0.081	0.6267
2569693	41213	10245113	C/T	0.297	0.355	0.0389
281439	41419	10245319	G/C	0.526	0.589	0.0352
281440	41613	10245513	G/A	0.736	0.746	0.7085
ICAM SNPI	42407	10246307	C/G	0.325	0.394	0.0173
1333881	43440	10247340	T/C	0.336	0.360	0.3961
1056538	44247	10248147	T/C	0.592	0.489	0.0009
2228615	44677	10248577	A/G	0.595	0.519	0.0112
2569702	45256	10249156	T/C	0.294	0.357	0.0254
2569703	45536	10249436	C/G	0.438	0.476	0.2109
ICAM SNPJ	46153	10250053	C/T	Not Allelotyped		
2569707	47546	10251446	C/G	0.829	0.840	0.6238
2916060	47697	10251597	A/C	0.010	0.002	0.0702
885743	47944	10251844	A/T	Not Allelotyped		
ICAM SNPK	48530	10252430	C/G	Not Allelotyped		
892188	51102	10255002	T/C	0.512	0.434	0.0104
2291473	57090	10260990	T/C	0.087	0.090	0.8770
281416	60093	10263993	A/G	0.546	0.505	0.1669
281417	60439	10264339	T/C	0.471	0.476	0.8531
281418	62694	10266594	G/C	0.914	0.934	0.1968
430092	66260	10270160	C/T	0.229	0.257	0.2758
368835	67295	10271195	A/G	0.703	0.727	0.3808
2358583	67304	10271204	T/G	0.304	0.326	0.4322
ICAM SNPL	67731	10271631	G/T	0.705	0.669	0.2029
1045384	68555	10272455	C/A	0.180	0.187	0.7736
281427	70429	10274329	C/T	0.217	0.176	0.0916
3745264	70875	10274775	T/G	0.853	0.836	0.4285
281426	72360	10276260	G/A	0.565	0.685	0.0001
281424	74228	10278128	C/T	0.246	0.250	0.8929
281423	76802	10280702	C/T	0.192	0.197	0.8585



dbSNP rs#	Position in Figure 1	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
281422	77664	10281564	T/C	0.632	0.632	0.9791
281421	78803	10282703	A/G	0.920	0.925	0.7863
281420	79263	10283163	A/G	0.392	0.432	0.1774
3745263	80810	10284710	A/G	0.936	0.923	0.4005
3745261	81020	10284920	T/C	0.006	0.008	0.5979
3181049	82426	10286326	T/C	0.650	0.640	0.7183
281412	82783	10286683	T/C	0.408	0.352	0.0527
2230399	85912	10289812	C/G	0.826	0.838	0.5900
2278442	86135	10290035	G/A	0.581	0.594	0.6511
2304237	87877	10291777	T/C	0.102	0.093	0.6063
281413	88043	10291943	G/A	Not Allelotyped		
1058154	88206	10292106	A/C	0.780	0.810	0.2203
3176769	88343	10292243	T/C	0.199	0.214	0.5539
2304240	90701	10294601	G/A	0.170	0.203	0.1661
3176768	90974	10294874	A/G	0.642	0.650	0.7681
3176767	91060	10294960	C/A	0.727	0.725	0.9511
3176766	91087	10294987	C/T	0.230	0.231	0.9513
ICAM_SNP	91594	10295494	G/A	0.289	0.267	0.4128
281415	92302	10296202	T/G	0.754	0.766	0.6399
3176764	92384	10296284	A/G	0.899	0.894	0.8086
281412	NOT MAPPED			0.154	0.156	0.9342
281413	NOT MAPPED			0.299	0.302	0.9195
281415	NOT MAPPED			0.664	0.684	0.4825

[0270] Figure 14 shows the proximal SNPs in and around the *ICAM* region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 14 can be determined by consulting Table 12. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0271] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, e.g., see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to

the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than  $10^{-8}$  were truncated at that value.

[0272] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link ([http address: www.ncbi.nlm.nih.gov/LocusLink/](http://www.ncbi.nlm.nih.gov/LocusLink/)) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is placed at the 3' end of each gene to show the direction of transcription.

#### Additional Genotyping

[0273] In addition to the ICAM region incident SNP, two other SNPs were genotyped in the discovery cohort. The discovery cohort is described in Example 1. The SNPs (rs1801714 and rs2228615) are located in the ICAM5 encoding portion of the sequence, were associated with breast cancer with a p-value of 0.0734 and 0.00236, respectively, and encoded non-synonymous amino acids (see Table 15).

[0274] The methods used to verify and genotype the two proximal SNPs of Table 15 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 13 and Table 14, respectively.

**Table 13**

dbSNP rs#	Second PCR primer	First PCR primer
1801714	ACGTTGGATGAGGGTTGCAGAGCAGGAGAA	ACGTTGGATGAGCCAAGGTGACGCTGAATG
2228615	ACGTTGGATGAGATGGTGACAGTAACCTGC	ACGTTGGATGTGGCATTAGCTGAAGCTGG

**Table 14**

dbSNP rs#	Extend Primer	Term Mix
1801714	CCTTCAGCAGGAGCTGGGCCCTC	ACT
2228615	TAACCTGCGCAGCTGGG	ACT

[0275] Table 15, below, shows the case and control allele frequencies along with the p-values for the SNPs genotyped. The disease associated allele of column 4 is in bold and the disease associated amino acid of column 5 is also in bold. The chromosome positions provided correspond to NCBI's Build 33.

**Table 15: Genotyping Results**

dbSNP rs#	Position in Figure 1	Chrom - some Position	Alleles (A1/A2)	Amino Acid Change	AF F case	AF F control	p-value	Odds Ratio
1801714	36517	10240417	T/C	L352P	T = 0.010 C = 0.990	T = 0.030 C = 0.097	<b>0.0734</b>	2.260
2228615	44677	10248577	A/G	T348A	A = 0.340 G = 0.660	A = 0.430 G = 0.570	<b>0.00236</b>	1.470

Example 5

MAPK10 Proximal SNPs

[0276] It has been discovered that a polymorphic variation (rs1541998) in a region that encodes MAPK10 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs1541998) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately sixty-three allelic variants located within the MAPK10 region were identified and allelotyped. The polymorphic variants are set forth in Table 16. The chromosome position provided in column four of Table 16 is based on Genome “Build 33” of NCBI’s GenBank.

**Table 16**

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
2575681	4	191	87306691	C/T
2575680	4	1490	87307990	A/G
2589505	4	3781	87310281	C/T
2589504	4	3935	87310435	G/A
2164538	4	4512	87311012	T/C
2575679	4	7573	87314073	A/G
MAP_SNP1	4	8467	87314967	A/T
2869408	4	9001	87315501	C/G
934648	4	9732	87316232	T/C
2164537	4	13477	87319977	T/C
2575678	4	13787	87320287	A/C
2575677	4	13903	87320403	G/C
2589509	4	14355	87320855	T/G
2164536	4	15053	87321553	A/C
2164535	4	15459	87321959	T/A
MAP_SNP2	4	17762	87324262	G/A
2589523	4	19482	87325982	C/T
3755970	4	19631	87326131	A/C

dbSNP rs#	Chromosome	Position in Figure 2	Chromosome Position	Allele Variants
2575675	4	22170	87328670	G/A
1202	4	22688	87329188	T/C
1201	4	22748	87329248	A/G
2589516	4	23376	87329876	G/T
2575674	4	23826	87330326	A/T
2589515	4	23868	87330368	G/C
MAP_SNP3	4	24154	87330654	C/T
2589506	4	25972	87332472	G/A
1436524	4	26057	87332557	A/G
2575672	4	26361	87332861	C/T
2589518	4	26599	87333099	G/A
3775164	4	26712	87333212	T/G
2589514	4	26812	87333312	G/A
3775166	4	27069	87333569	T/C
3775167	4	32421	87338921	C/T
3775169	4	33557	87340057	T/C
2043650	4	35127	87341627	A/G
2043649	4	35222	87341722	T/G
3775170	4	35999	87342499	T/A
1541998	4	36424	87342924	C/T
2043648	4	37403	87343903	A/G
2282598	4	39203	87345703	C/T
2282597	4	39226	87345726	G/A
3775173	4	41147	87347647	T/C
1469870	4	46176	87352676	G/C
1436522	4	50452	87356952	T/C
1946733	4	52919	87359419	G/A
1436525	4	60214	87366714	G/A
3822037	4	61093	87367593	C/G
3775176	4	62572	87369072	G/A
1436527	4	63601	87370101	C/T
1436529	4	65362	87371862	T/C
3775182	4	65863	87372363	T/G
3775183	4	66207	87372707	G/A
3775184	4	66339	87372839	A/G
3775187	4	69512	87376012	T/C
1010778	4	70759	87377259	A/G
2282596	4	71217	87377717	T/A
2118044	4	73382	87379882	A/T
1469869	4	76307	87382807	C/T
1046706	4	Not mapped		G/T
2060588	4	Not mapped		G/A
2289490	4	Not mapped		C/T
2289491	4	Not mapped		C/T
729511	4	Not mapped		T/C

### Assay for Verifying and Allelotyping SNPs

[0277] The methods used to verify and allelotype the proximal SNPs of Table 16 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 17 and Table 18, respectively.

**Table 17**

dbSNP rs#	Forward PCR primer	Reverse PCR primer
958	ACGTTGGATGATCCGCATGTGTCTGTATTG	ACGTTGGATGCCAGTGCATTATGTCTTGG
1201	ACGTTGGATGTGCCAGTGCTCTGAAACTG	ACGTTGGATGCCTGTGGTCTCTATTGCTTG
1201	ACGTTGGATGACAAGAATGCCAGTGCTCTG	ACGTTGGATGCCTGTGGTCTCTATTGCTTG
1202	ACGTTGGATGTAATCTCAGAATGGCAGCAC	ACGTTGGATGTCAAGCAATAGAGACCACAG
10305	ACGTTGGATGTTCAAGAATTATTTATTGCAA GTC	ACGTTGGATGGGTGAAGCTTGAAAGCAAGC
729511	ACGTTGGATGTTAATGTAGTAAAAAGCACG	ACGTTGGATGCTAGAGATCGGTTTTACACC
934648	ACGTTGGATGACTGGTTGATACCATAGGAC	ACGTTGGATGTGTACTGCTTTCATCCTTGC
934648	ACGTTGGATGACTGGTTGATACCATAGGAC	ACGTTGGATGTGTACTGCTTTCATCCTTGC
1010778	ACGTTGGATGCAGAGGAAAGAAAAGTAAAG	ACGTTGGATGGGATTTGTTCTTAATCTTTC
1046706	ACGTTGGATGCAATGGGAGTCAAGTCCTC	ACGTTGGATGTTTTGCTCCTAAGCTGAAGG
1436522	ACGTTGGATGGGAATTGAAATTGGCATTGC	ACGTTGGATGATTGGAAGGAGGAAGCATAG
1436524	ACGTTGGATGGAGTTGCCAGTAGCTTTGAG	ACGTTGGATGATTGTTCCAGGGTGCTCTG
1436525	ACGTTGGATGGTGCAATCTTGTTCACTGC	ACGTTGGATGGCTTACACTAGCTACTTGGG
1436527	ACGTTGGATGAGCACTGTGAGTTAAACCTG	ACGTTGGATGCTGTATAGAGAGCTGTTTGC
1436529	ACGTTGGATGCTATGGCAGCAGAAGAGTAG	ACGTTGGATGAATGTTGGACCACATGTACG
1469869	ACGTTGGATGCATGGCGAGGAAATCTGTTT	ACGTTGGATGTTGATATATCAGAGCCTTG
1469870	ACGTTGGATGATACTGAGCTCCATTTTGGG	ACGTTGGATGATGGCACAGTTTAGCATGTC
1541998	ACGTTGGATGGCCCATGTTAACATTTTCTTC	ACGTTGGATGCTGATTATTCTGATGGTAATG
1946733	ACGTTGGATGGCAGGAGGATAGATCTGTAG	ACGTTGGATGTAGCTTCTAAACATCTCTTG
2043648	ACGTTGGATGTGGCTTTCTGAATGCTAGAG	ACGTTGGATGAGGGCGGAATGATTTTATGC
2043649	ACGTTGGATGGCACTACATGGGACACAAAG	ACGTTGGATGGTCTACTAGTCCCTGTATG
2043650	ACGTTGGATGGCTGAGGGAGAAATTGAGTG	ACGTTGGATGCTGTGCCTTGACATAGTAG
2060588	ACGTTGGATGTTTCATTGCTCATGGATTAG	ACGTTGGATGGATAAGTATTGGCTTAATCTG
2118044	ACGTTGGATGAACAACCTTGCTAATTCTAC	ACGTTGGATGGTCATTGCCTCTAGCTAGTG
2164535	ACGTTGGATGACCAGCACTATTACCCATGC	ACGTTGGATGGAATGATGTAAACGTTGGAG
2164536	ACGTTGGATGGTGATGAAAACCATGTGAGC	ACGTTGGATGCTGGAGAACAAAAGACCACC
2164537	ACGTTGGATGCAAGGCCAAAATGTTTCCAGC	ACGTTGGATGAACACACTTAGTACCCACGC
2164538	ACGTTGGATGTACTGCAGAGCTCTCCCTTG	ACGTTGGATGAGAGGTCATCTTAATGGGCC
2282596	ACGTTGGATGTCATACTGATCAACCTGAAG	ACGTTGGATGGGTGGCTTTGTGAAACCTTG
2282597	ACGTTGGATGGCATGGTTCTGTTATAAGGC	ACGTTGGATGACACTTGATTACAATGGCCC
2282598	ACGTTGGATGCACGCCTAAGCAATTAATGAC	ACGTTGGATGGTGAATGAAGGAAAAGTAGC
2289490	ACGTTGGATGTGATTACTGGATTGGCTGGG	ACGTTGGATGAAATGCCCTGAAGACCCAGC
2289491	ACGTTGGATGGGAATGCATTGTAAACCAGG	ACGTTGGATGACCTAGCCTTGACAGGAGAC
2575672	ACGTTGGATGATAGTGTTATCATAGAGACC	ACGTTGGATGCTCCAGGAGCAAGGATTATG
2575674	ACGTTGGATGGTGGGTAACAGTTTTCAGGC	ACGTTGGATGCTCTCCTACTCTTTACTGTC
2575675	ACGTTGGATGTCGTACCTGCATAAGTGGTG	ACGTTGGATGTTGGGAAGGTACTAACAGCG
2575677	ACGTTGGATGGATGCCAATTTGGTTTGGCC	ACGTTGGATGGAAGGATAAGCCACAGTGAG
2575678	ACGTTGGATGCTTCAAGAGGCCATACAGAC	ACGTTGGATGAAGCACCATTGTGGCTCAG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
2575679	ACGTTGGATGCTTTCCTGCTGCATTTAGTG	ACGTTGGATGTAAGCCAGTAACACATGCCG
2575680	ACGTTGGATGGCCCTGAAGTTTTTGAATGG	ACGTTGGATGGAGCCCAATACAATCAGGTG
2575681	ACGTTGGATGTTCACTGCTAACATGCATGG	ACGTTGGATGTTATATAGCCTTCTTTTCTC
2589504	ACGTTGGATGGGATAGGAAACATATTAAGG	ACGTTGGATGCTGTGTGATTTGGACAACCC
2589505	ACGTTGGATGAGACTGTAGCCTAAATGAGG	ACGTTGGATGCATTTTATGAGAAGATGCAC
2589506	ACGTTGGATGGCAACTCAGCTAGCCTTTAC	ACGTTGGATGTGTTATGCGGGAGTATAAGG
2589509	ACGTTGGATGTGAATCATGGTTGCCTCCTG	ACGTTGGATGATACGCAGGTTGTAGAGAGG
2589514	ACGTTGGATGTATACATTGTCCTGATAGAG	ACGTTGGATGCTTAAATGTCTCTAGAAAAGG
2589515	ACGTTGGATGCACCTGTATACCAATTTGTAG	ACGTTGGATGGCCAAACCATTTTGTGCCTG
2589516	ACGTTGGATGCATACTCTGCCAAAGTTTTA	ACGTTGGATGACTCACACTGTGGTTTGGGG
2589518	ACGTTGGATGCCAGGCAAAAAGAATGACCG	ACGTTGGATGAATGATATGCACCGATCTTC
2589523	ACGTTGGATGTCATGTAGCTAAACAAAGGC	ACGTTGGATGAGCAGGGTTAAATTTCCCAG
2589525	ACGTTGGATGAAGAACATTGAAAGAAGCAG	ACGTTGGATGGTATTTAAATTAGTGGTGTG
2869408	ACGTTGGATGTCCCAGTACCTAAGTAGCAG	ACGTTGGATGGCTTTGAATTACTCTGTCCC
3755970	ACGTTGGATGTACAACTAGTATCTACAGAC	ACGTTGGATGGTGACCATGTAGAAATCTGTG
3775164	ACGTTGGATGGAACATGAAAAATTCATAAGC	ACGTTGGATGAAGTTTCCCTGGTCGTGATC
3775166	ACGTTGGATGCTGTTTTTCACCCCCGATTC	ACGTTGGATGCTGAGGAGTCCATCATAGTG
3775167	ACGTTGGATGGAAACAAGCAGATGTCATGG	ACGTTGGATGGCTTCTGATTTTATATGGCAC
3775169	ACGTTGGATGGGGAGAGAATGGTTGCATAT	ACGTTGGATGATGCTGAACAACAGGATGGG
3775170	ACGTTGGATGCCTAAGACCTATGCTCTCAC	ACGTTGGATGCCATTTTTTGCTAGCAGGAG
3775173	ACGTTGGATGCAAGAGGGCTGCTTTAAACC	ACGTTGGATGTAAATTTGCAGAGGCCGTCG
3775176	ACGTTGGATGAAAAGGTCACCAAGTGACCTG	ACGTTGGATGTAGTCCAAGTATTTCCCAAG
3775182	ACGTTGGATGGATATCTCCCTCCTATTGGC	ACGTTGGATGGCTGGACTCTATTAGGCCAT
3775183	ACGTTGGATGGATCTCTGATCTTAGACCAC	ACGTTGGATGTGCAGATATGTAGGCCAAGC
3775184	ACGTTGGATGGACCAGCAACCATGATGAAG	ACGTTGGATGGTTCTACTTTGACCACAGGC
3775187	ACGTTGGATGTAGCACCTTCAGGATCTTTC	ACGTTGGATGAATCATGATCCCAGGGCAAG
3822037	ACGTTGGATGGTAATCCATAAACTGTGGGAG	ACGTTGGATGTCCCACCCTGACTTCTTTGC

Table 18

dbSNP rs#	Extend Primer	Term Mix
958	TTATGTCTTGGTAGAGCC	ACG
1201	TCTATTGCTTGAAGAGAGAAAAG	ACT
1201	TTGCTTGAAGAGAGAAAAG	ACT
1202	CCACCTGCACCATCGCCAT	ACT
10305	AGCTAAATTGCAACAACA	ACG
729511	ATTGAACTGTATACTTAAAAATGC	ACT
934648	ACTCTCCCACTGAGCAAGC	ACT
934648	ACTCTCCCACTGAGCAAGC	ACT
1010778	TTGAAATACTGTTTGTTCCTCAA	ACT
1046706	TCCTAAGCTGAAGGGAATGC	CGT
1436522	GAGGAAGCATAGATTTGGTGT	ACT
1436524	CCAGGGTGCTCTGGTTTAATT	ACT
1436525	GGCTTAAACCTGGGAGG	ACG
1436527	GAGCTGTTTGCATTTATAACTCA	ACG

dbSNP rs#	Extend Primer	Term Mix
1436529	ACCACATGTACGTAAGGGGA	ACT
1469869	AAACACCATCTACTCTGAAGAA	ACG
1469870	CTTATATTCTCTGTGGCACCAA	ACT
1541998	ATTATTCTGATGGTAATGATCCAG	ACG
1946733	CTAAACATCTCTTGAATATTCTG	ACG
2043648	TGATTTTGTAGCTAAAGGGGACA	ACT
2043649	CCTCTTGTCTTATTATCCC	ACT
2043650	GCACATAGTAGTAGCTCA	ACT
2060588	ATTGGCTTAATCTGTACATCAATT	ACG
2118044	GTGGGGTTAGATATTATTTCTGA	CGT
2164535	GATAAATGTGAGATTGAGAGA	CGT
2164536	CCTGTGTTCTTTGTATTTATAT	ACT
2164537	CGGCTTCTACTCTCTTATTCA	ACT
2164538	GTCACATTCTTACCCTC	ACT
2282596	GAAACCTTGCATGAACT	CGT
2282597	CAGAAGCTACTTTTCCTTCA	ACG
2282598	AGGAAAAGTAGCTTCTGGG	ACG
2289490	GCTAGACTCCTGATACC	ACG
2289491	GGCTTGCTCCTGGTAATTTA	ACG
2575672	CAAGGATTATGTTAACCCT	ACG
2575674	TATTCACACCTGCCTTC	CGT
2575675	GTTCTTGCCTGGTTTAC	ACG
2575677	GGAATGAGGGCAACAGGA	ACT
2575678	TGTGGCTCAGGTCCAGG	ACT
2575679	CTTCTGGACATTAAATTGT	ACT
2575680	GGATGCATGGTTTCTCTAAT	ACT
2575681	TTCTTTTCTCTTTTAGGAATCT	ACG
2589504	GTGCTAGGATCCTCAGT	ACG
2589505	GTTTTAGCATAATTGCTTCTTTA	ACG
2589506	GAGAAGAAACCTGCCCA	ACG
2589509	AGGGCTGCAGGGAAGAT	ACT
2589514	AGAAAAGGTTTTTAAAGTCCTC	ACG
2589515	GAAACTGTTACCCACTC	ACT
2589516	GGTTTGGGGTTTCATT	CGT
2589518	TGCACCGATCTTCAAATAAA	ACG
2589523	TTTCCCAGATTAATTATCAGATT	ACG
2589525	TTAGTGGTGTGACTTGCA	ACG
2869408	CGAATCTCTTTAACTGCTG	ACT
3755970	GGTTTCTTCTAAACTGACCT	ACT
3775164	TTTTTTGGGATCTTGATATTTTA	ACT
3775166	AACTTATGAAAGAATATGAAGGAT	ACT
3775167	TAAGAGAAGTCTTCAGTGCTT	ACG
3775169	GCAGAGATTTTTCAAATCTCTAA	ACT
3775170	TTTTTAAAGCTGAAAATAAACCA	CGT

dbSNP rs#	Extend Primer	Term Mix
3775173	GCCGTCGAACAAATACT	ACT
3775176	TATTTCCCAAGTGCCCA	ACG
3775182	CTGTCAGTTGCCTTAGG	ACT
3775183	AGTCAAGACCAGCTGGG	ACG
3775184	CTCTTTCTTCTGATCCC	ACT
3775187	AGTGCATTACAGTGGTC	ACT
3822037	TTTGCTTATTTTCATAGAAGGAAT	ACT

### Genetic Analysis of Allelotyping Results

[0278] Allelotyping results are shown for cases and controls in Table 19. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where “AF” is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 ( $A1\ AF = 1 - A2\ AF$ ). For example, the SNP rs2575681 has the following case and control allele frequencies: case A1 (C) = 0.611; case A2 (T) = 0.389; control A1 (C) = 0.632; and control A2 (T) = 0.368, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

**Table 19**

dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2575681	191	87306691	C/T	0.389	0.368	0.483
2575680	1490	87307990	A/G	0.599	0.585	0.646
2589505	3781	87310281	C/T	0.484	0.493	0.753
2589504	3935	87310435	G/A	0.258	0.274	0.563
2164538	4512	87311012	T/C	0.403	0.412	0.784
2575679	7573	87314073	A/G	0.020	0.003	0.006
MAP_SNP1	8467	87314967	A/T	0.704	0.682	0.441
2869408	9001	87315501	C/G	0.708	0.716	0.777
934648	9732	87316232	T/C	0.655	0.664	0.741
2164537	13477	87319977	T/C	0.262	0.306	0.109
2575678	13787	87320287	A/C	0.110	0.078	0.065
2575677	13903	87320403	G/C	0.920	0.991	0.000
2589509	14355	87320855	T/G	0.198	0.209	0.668
2164536	15053	87321553	A/C	0.623	0.605	0.534
2164535	15459	87321959	T/A	0.573	0.571	0.944
MAP_SNP2	17762	87324262	G/A	0.389	0.401	0.693
2589523	19482	87325982	C/T	0.779	0.813	0.156
3755970	19631	87326131	A/C	0.118	0.107	0.563
2575675	22170	87328670	G/A	0.656	0.694	0.176
1202	22688	87329188	T/C	0.764	0.762	0.933
1201	22748	87329248	A/G	0.128	0.117	0.579
2589516	23376	87329876	G/T	0.427	0.478	0.086
2575674	23826	87330326	A/T	0.583	0.666	0.004
2589515	23868	87330368	G/C	0.413	0.461	0.106



dbSNP rs#	Position in Figure 2	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
MAP_SNP3	24154	87330654	C/T	0.175	0.158	0.430
2589506	25972	87332472	G/A	0.435	0.491	0.063
1436524	26057	87332557	A/G	0.660	0.756	0.001
2575672	26361	87332861	C/T	0.274	0.185	0.001
2589518	26599	87333099	G/A	0.194	0.130	0.004
3775164	26712	87333212	T/G	0.073	0.080	0.644
2589514	26812	87333312	G/A	0.445	0.358	0.004
3775166	27069	87333569	T/C	0.249	0.167	0.001
3775167	32421	87338921	C/T	0.156	0.152	0.882
3775169	33557	87340057	T/C	0.169	0.130	0.067
2043650	35127	87341627	A/G	0.697	0.787	0.001
2043649	35222	87341722	T/G	0.698	0.763	0.016
3775170	35999	87342499	T/A	0.207	0.220	0.596
1541998	36424	87342924	C/T	0.715	0.772	0.029
2043648	37403	87343903	A/G	0.424	0.466	0.159
2282598	39203	87345703	C/T	0.022	0.031	0.324
2282597	39226	87345726	G/A	0.817	0.802	0.541
3775173	41147	87347647	T/C	0.158	0.148	0.645
1469870	46176	87352676	G/C	0.118	0.063	0.002
1436522	50452	87356952	T/C	0.165	0.120	0.036
1946733	52919	87359419	G/A	0.240	0.226	0.588
1436525	60214	87366714	G/A	0.054	0.039	0.212
3822037	61093	87367593	C/G	0.956	0.918	0.010
3775176	62572	87369072	G/A	0.969	0.909	0.000
1436527	63601	87370101	C/T	0.288	0.251	0.175
1436529	65362	87371862	T/C	0.555	0.534	0.481
3775182	65863	87372363	T/G	0.858	0.870	0.568
3775183	66207	87372707	G/A	0.565	0.617	0.080
3775184	66339	87372839	A/G	0.174	0.185	0.634
3775187	69512	87376012	T/C	0.307	0.291	0.575
1010778	70759	87377259	A/G	0.330	0.275	0.048
2282596	71217	87377717	T/A	0.735	0.738	0.892
2118044	73382	87379882	A/T	0.352	0.319	0.248
1469869	76307	87382807	C/T	0.388	0.335	0.069
1046706	Not mapped		G/T	0.538	0.533	0.866
2060588	Not mapped		G/A	0.188	0.135	0.016
2289490	Not mapped		C/T	0.780	0.812	0.187
2289491	Not mapped		C/T	0.960	0.971	0.297
729511	Not mapped		T/C	0.864	0.866	0.914

[0279] Figure 15 shows the proximal SNPs in and around the MAPK10 region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 15 can be determined by consulting Table 19. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0280] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to

expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than  $10^{-8}$  were truncated at that value.

[0281] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link ([http address: www.ncbi.nlm.nih.gov/LocusLink/](http://www.ncbi.nlm.nih.gov/LocusLink/)) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is placed at the 3' end of each gene to show the direction of transcription.

#### Example 6

##### KIAA0861 Proximal SNPs

[0282] It has been discovered that a polymorphic variation (rs2001449) in a gene encoding KIAA0861 is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs2001449) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. A total of sixty-three allelic variants located within or nearby the KIAA0861 gene were identified and fifty-severn allelic variants were allelotyped. The polymorphic variants are set forth in Table 20. The chromosome position provided in column four of Table 20 is based on Genome "Build 33" of NCBI's GenBank.

**Table 20**

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
3811729	3	107	184282507	A/G
693208	3	2157	184284557	C/G
488277	3	7300	184289700	T/C
645039	3	8233	184290633	T/C
670232	3	9647	184292047	A/T
575326	3	9868	184292268	T/C
575386	3	9889	184292289	C/G
471365	3	10621	184293021	G/C

dbSNP rs#	Chromosome	Position in Figure 3	Chromosome Position	Allele Variants
496251	3	11003	184293403	G/A
831246	3	11507	184293907	T/C
831247	3	11527	184293927	G/C
831249	3	11718	184294118	C/T
831250	3	11808	184294208	T/C
831252	3	12024	184294424	T/C
512071	3	13963	184296363	C/T
1502761	3	14300	184296700	A/C
681516	3	14361	184296761	C/T
619424	3	16287	184298687	T/G
529055	3	18635	184301035	A/G
664010	3	19365	184301765	T/G
2653845	3	24953	184307353	G/A
472795	3	25435	184307835	G/A
507079	3	26847	184309247	G/A
534333	3	27492	184309892	T/C
831242	3	27620	184310020	T/C
536111	3	27678	184310078	C/T
536213	3	27714	184310114	G/A
831245	3	29719	184312119	A/G
639690	3	30234	184312634	T/C
684174	3	31909	184314309	T/C
571761	3	32153	184314553	C/G
1983421	3	33572	184315972	T/C
2314415	3	42164	184324564	T/G
2103062	3	43925	184326325	A/G
6804951	3	45031	184327431	C/T
1403452	3	45655	184328055	T/C
903950	3	48350	184330750	C/A
2017340	3	48418	184330818	A/G
2001449	3	48563	184330963	G/C
3821522	3	53189	184335589	A/G
1390831	3	56468	184338868	T/G
1353566	3	59358	184341758	C/A
1813856	3	63761	184346161	C/T
2272115	3	65931	184348331	G/A
3732603	3	67040	184349440	G/C
940055	3	69491	184351891	A/C
2314730	3	83308	184365708	A/G
KIAA0861_373 2602	3	126545	184408945	C/T
KIAA0861_229 3203	3	137592	184419992	A/T
7639705	3	147169	184429569	G/T

### Assay for Verifying and Allelotyping SNPs

[0283] The methods used to verify and allelotype the sixty-three proximal SNPs of Table 20 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 21 and Table 22, respectively.

**Table 21**

dbSNP rs#	Forward PCR primer	Reverse PCR primer
471365	ACGTTGGATGTGAGTGACATTTGTGTCACC	ACGTTGGATGCGGAGGATCTGAACAACCTTC
472795	ACGTTGGATGTCACCTGAGCATCAGACATG	ACGTTGGATGATAGTGGAAGGAGAAACGGG
484315	ACGTTGGATGGTTCTAATGTCACCCCTTCC	ACGTTGGATGCAATGTGGCAAATTCTCTGG
488277	ACGTTGGATGCACACATTCTTCTCAAGTGC	ACGTTGGATGGGAGGGACACAATTTAACTC
496251	ACGTTGGATGGGGAGTCATTCCAATACCAG	ACGTTGGATGGGAGTGAAAGGTCATATTGG
502289	ACGTTGGATGATCACTGCAACCTCCACCTC	ACGTTGGATGTGTGGCATGAGCCTGTAATC
507079	ACGTTGGATGAAGCCTCAGATGAGGCATAC	ACGTTGGATGTCTGAAAGGGTTCAGGAAGG
512071	ACGTTGGATGCAAATCACCCCTGACAATTC	ACGTTGGATGACCAGCACACTCAGCTTTAG
519088	ACGTTGGATGTCACCTGAGGTCAGGAGTTG	ACGTTGGATGAGGTTTCACCATGTTAGCCG
529055	ACGTTGGATGCTGCAGTTATCTGGGTGAGC	ACGTTGGATGCCAGAACGTGGCTTGTGGG
534333	ACGTTGGATGCGTTGATGCACTGAAGGGAG	ACGTTGGATGAGAGGCTAAATGTTGGCAGG
536111	ACGTTGGATGTGTATCTGATCCCAGGTCAC	ACGTTGGATGATTGGTGTTAAGTGGCGTGC
536213	ACGTTGGATGTGAGGACCTCATTATTGGTG	ACGTTGGATGCTGAGCAATCGAACTGCTAC
571761	ACGTTGGATGAATATCCTAGGCTAGCAGTG	ACGTTGGATGGTGCATAAATACATGAATAG
575326	ACGTTGGATGACAGAGAGGCTTGGTCATAC	ACGTTGGATGGGTGCTTGGTTGTGATTCTC
575386	ACGTTGGATGATTCCCTGCAGGTACTGTGTC	ACGTTGGATGTGAGCCCAAACTACTGCTG
578886	ACGTTGGATGATGAAGTCTCGCTCTGTTGC	ACGTTGGATGAATCACTTGAACCCAGGAGG
602646	ACGTTGGATGTCTGGGACCGTTTACCGCA	ACGTTGGATGGAGGAGACCCAGGGTATGAG
619424	ACGTTGGATGACCGGGAGCTCCCAGTCTG	ACGTTGGATGTGGGAATCGGTTGAGAGCCG
620722	ACGTTGGATGTAAGGCGCCTGCAGAGGCCA	ACGTTGGATGGCAGCAAAGAATTGCCCGGC
631755	ACGTTGGATGATTTGTAGCTTTGCCCCAGC	ACGTTGGATGTTTGTGAGCTCCAAGTTGGG
639690	ACGTTGGATGGCATTTTACCACCATGTGGTT	ACGTTGGATGCCTTCATGTTAATTCTGCCC
645039	ACGTTGGATGCCTCTGAGTTCCTCAGTTT	ACGTTGGATGTTATCACCCCTGCTGTCCTAC
664010	ACGTTGGATGTGGTACCTCCAGGTAAAATG	ACGTTGGATGTCCAGGCAGTCATTTTACCC
670232	ACGTTGGATGGAAGGTGGAGCAGACATTAG	ACGTTGGATGACCTTAGTTATACCAGGCAC
678454	ACGTTGGATGTTAAGCCAGTCCCCACAAGG	ACGTTGGATGTTCTCTGCGGAGGAAAGTGC
681516	ACGTTGGATGCTCCTCCTCAGAGGACTAAC	ACGTTGGATGAGCCCAAGGACTCATACAAC
683302	ACGTTGGATGACCACGCTGGCTAATTTTG	ACGTTGGATGAAACATGGCGAAACCCGGTC
684174	ACGTTGGATGCTTTACTGAGTGGGCAAACG	ACGTTGGATGTCTAAGTGGAAGTCAAGCAGC
684846	ACGTTGGATGAAGTTCCTCTGGTGGACAAC	ACGTTGGATGACCACCAGATAAAATCCCTC
693208	ACGTTGGATGTTTTGACAGGGCTTGAGTCC	ACGTTGGATGGCTGAAAGCCCTCAATCTAG
831242	ACGTTGGATGCAATTGCTCAGACCTTCACC	ACGTTGGATGAATGCTAGAGACATTGCACC
831245	ACGTTGGATGCTAGAATTACAGGTGCACAC	ACGTTGGATGGCCAAGATGGTGAAACCTTG
831246	ACGTTGGATGCACAATCTGTTAGAATGGTGG	ACGTTGGATGCGTCAAGACTGAATGCATAG
831247	ACGTTGGATGGAAAATATAGTCCTACACAA	ACGTTGGATGCGTCAAGACTGAATGCATAG
831249	ACGTTGGATGTCTCCTAATGCTATCCCTCC	ACGTTGGATGAACACATGGACACAGGAAGG
831250	ACGTTGGATGAGGGACATGGATGAAATTGG	ACGTTGGATGAATCCCACCTATGAGTGAG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
831252	ACGTTGGATGTGGGTATATACCCAAAGGAC	ACGTTGGATGGGTGGTTCCAAGTCTTTGC
903950	ACGTTGGATGCTTCAGTTCAGGGAGAGATC	ACGTTGGATGATAGGGCCCCCAGCATAAAA
940054	ACGTTGGATGTGGTAGAGATGAGGTCTTGC	ACGTTGGATGAAAGGCAGGAGGATTGCTTG
940055	ACGTTGGATGTATGCTTCCAGTCTCTGACC	ACGTTGGATGATAGGTAATCCAGTTGGGCC
1353566	ACGTTGGATGGGTGTACTCTGCCATTTGTC	ACGTTGGATGTGGAGGAGGTTCTAGTACCC
1390831	ACGTTGGATGGTCTGCCAAAGTTCCCTTAG	ACGTTGGATGAGGAAAGGGAAGAGAAACCG
1403452	ACGTTGGATGCAGAAGTTAGGATGCAGATG	ACGTTGGATGCCAGTAGAGATAGAATTTTGG
1502761	ACGTTGGATGCAGAAATATGAAGGTGGCCC	ACGTTGGATGACCTTGAGCTCTGAGCCCTT
1629673	ACGTTGGATGAAGGATCACGTGAAGTCAGG	ACGTTGGATGGGCACCATGTGTGGCTAATT
1813856	ACGTTGGATGTCTGACTCCCTGATTCAAGC	ACGTTGGATGACAAAAATTAGCCGGGCGTG
1983421	ACGTTGGATGTCCAGGTGTTATGGAGTCAG	ACGTTGGATGGGCTTCTTGTGCTGCTGTGT
2001449	ACGTTGGATGATGTCAAGTGCACCCACATG	ACGTTGGATGAGGAAGAAACTGACGGAAGG
2017340	ACGTTGGATGTATTCCACTGCCTGCTTTCC	ACGTTGGATGGAAAACAGGAGGAAGTGGTG
2030578	ACGTTGGATGTTCTCCACTTTCTGGTCAAC	ACGTTGGATGAACAACCTTACTTCATGCCC
2049280	ACGTTGGATGCTTCCCAACATTTTCGGCTC	ACGTTGGATGTGGATACTGAGGGTCAACTG
2103062	ACGTTGGATGTGCAGCCCTCAACCTTTCAG	ACGTTGGATGCCTTATTAGTTACTATTACG
2272115	ACGTTGGATGAGTTGTGAGTGATTTCAAGG	ACGTTGGATGCAGGCCCTTCTTGCTCTTATC
2272116	ACGTTGGATGATCTGTTGCCTTAGGTTTAC	ACGTTGGATGCTGTGCCTTCTGAGTAGTTC
2314415	ACGTTGGATGGGCTGAGTAACAGTCCATTG	ACGTTGGATGCTTACAGTATCCAAAAAGGG
2314730	ACGTTGGATGCTCAGGTAATCTGCCTTCTC	ACGTTGGATGCAGGGATAATGAGAACAAATC
2653845	ACGTTGGATGATCACTTGGACTCAGGAAGC	ACGTTGGATGAGTCTTGCTCTGTTTCCAGG
3732603	ACGTTGGATGCTCTCAATTCCATCAGTCTC	ACGTTGGATGCTTTACGAATTTCAACACAGG
3811728	ACGTTGGATGACGCGCCACACCTCCCTAC	ACGTTGGATGACGTGTCGGTCCCCCTTTCAT
3811729	ACGTTGGATGTGGGCGAGGTTCTGCAGCGT	ACGTTGGATGGTTTCGTTTCTCCGGCACAG
3811731	ACGTTGGATGTGCGGTAAACGGTCCCAGAG	ACGTTGGATGAACTCCGCCGGCCCCCTCCTA
3821522	ACGTTGGATGAACCCGCACTACAAGATTCC	ACGTTGGATGGTCAGTCCCACATTCAGAAC

Table 22

dbSNP rs#	Extend Primer	Term Mix
471365	TCCAAAACCACCAGATAAAATC	ACT
472795	GACATGTCCCTCTCGGCCT	ACG
484315	GGTATCAGGAAGAGTCA	ACT
488277	AGTGCACACAGAACATTTAACA	ACT
496251	GTATTGTCCTCCAGTGA	ACG
502289	CTGTAATCCCAGCTACTC	ACT
507079	GGCAATGTTTGCCCTTT	ACG
512071	CCCTGACAATTCCAAAATAA	ACG
519088	TTTCGCCATGTTTGCCAGG	ACG
529055	GAGCAGGCAGCACAAGT	ACT
534333	GGGAGAAAGTAACAGGGTC	ACT
536111	GTGAAGGTCTGAGCAAT	ACG
536213	TGGTGTAAAGTGGCGTG	ACG

dbSNP rs#	Extend Primer	Term Mix
571761	CTAGGCTAGCAGTGGGGTTG	ACT
575326	TGGTCATACCCTTCAAG	ACT
575386	GAAGGGTATGACCAAGC	ACT
578886	TGAGCCAAGATCATGCC	CGT
602646	CCAGGGTATGAGCGGAGGA	ACT
619424	TGCGGCCCGCCGGGTT	ACT
620722	GAATTGCCCGGCTCCGAAT	ACT
631755	TCCAAGTTGGGTCAAAG	ACT
639690	CTGCTATTCATTTGTGTAGA	ACT
645039	CCCTCAGTTTTTATTGATTATT	ACT
664010	ACCTCCAGGTAAAATGATTAGTT	ACT
670232	TGGGCAAACAAGCCCAT	CGT
678454	CAGGGATGGTAATTGAC	ACG
681516	GGCCACCTTCATATTTT	ACG
683302	CAGGAGATCCAGACCATCCC	ACG
684174	CTCTGATGTTACCTCCTCC	ACT
684846	AGTTGTTTCTGATCCTCC	ACT
693208	TCAATCTAGTGATAAGGAGGGT	ACT
831242	CAGGTGGATGGGGACAC	ACT
831245	CACACCACCACGCCCCGGCT	ACT
831246	AGAATGGTGGTGTATTTTTAC	ACT
831247	TAGTCCTACACAATCTGTTA	ACT
831249	GCTATCCCTCCCCCTTCCC	ACG
831250	GACAAAAAACCAACACC	ACT
831252	CTATAAAGACACATGCACAC	ACT
903950	AGATCACATTGCCAACCCCCA	CGT
940054	AAAGTAGCAGTTTGAGACCA	ACT
940055	GTCTCTGACCACTTGACCCA	ACT
1353566	TTGTCAGTTATGAGACCTTG	CGT
1390831	GGTTAGGAAGAAATCTGTG	ACT
1403452	CACAGATGCTCATGGGTCC	ACT
1502761	GGAGGAGGCACTATTAAT	ACT
1629673	TGTGGAGACAAGGTCTCACT	ACT
1813856	TCAAGCGATTCTCCTGC	ACG
1983421	GGCAGGGAAGAGAAGAGC	ACT
2001449	CACATGCCTGCTCGCCCCC	ACT
2017340	CCCTAAAGCATCTCACAGCCCC	ACT
2030578	TCATGCCCATTTGGGTTAG	ACT
2049280	GGGTCAACTGTACCAAG	ACG
2103062	GAGATCATTTCTCCTTCAAC	ACT
2272115	ATACCTCAGAATACAGCTTTTTTT	ACG
2272116	TCTCATTTCTCCTCTCTTTC	ACG
2314415	TAGTTGATGAAGATTGGG	ACT
2314730	TCCTTCTTCTCTGCTTT	ACT

dbSNP rs#	Extend Primer	Term Mix
2653845	AAGCGGAGGTTGCAGTGAGC	ACG
3732603	CTCATTTCCACCCTTCT	ACT
3811728	GTCCCCTTTTCATCTAAAC	ACT
3811729	TCTGCAGCGTGCGGCGA	ACT
3811731	CCTACCCCTACGGAGCC	ACT
3821522	GCATCTTCAGGAATCTTG	ACT

### Genetic Analysis of Allelotyping Results

[0284] Allelotyping results are shown for cases and controls in Table 23. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where “AF” is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP in row 2 of Table 13 (rs3811729) has the following case and control allele frequencies: case A1 (A) = 0.976; case A2 (G) = 0.024; control A1 (A) = 0.948; and control A2 (G) = 0.052, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped (“not AT”).

**Table 23**

dbSNP rs#	Position in Fig 3	Chrom Position	Alleles (A1/A2)	A2 Case AF	A2 Control AF	p-Value
3811729	107	184282507	A/G	0.024	0.052	<b>0.017</b>
693208	2157	184284557	C/G	0.186	0.207	0.368
3811731	not mapped		A/G	0.690	0.641	0.084
602646	not mapped		C/G	0.693	0.660	0.244
488277	7300	184289700	T/C	0.099	0.103	0.848
645039	8233	184290633	T/C	0.014	0.008	0.316
1629673	not mapped		T/C	0.064	0.093	0.069
670232	9647	184292047	A/T	0.865	0.863	0.932
575326	9868	184292268	T/C	0.128	0.129	0.949
575386	9889	184292289	C/G	0.776	0.779	0.905
684846	not mapped		C/G	0.799	0.745	0.033
471365	10621	184293021	G/C	0.746	0.740	0.815
496251	11003	184293403	G/A	0.156	0.160	0.853
831246	11507	184293907	T/C	0.773	0.802	0.243
831247	11527	184293927	G/C	0.829	0.826	0.879
831249	11718	184294118	C/T	0.071	0.051	0.160
831250	11808	184294208	T/C	0.682	0.697	0.589
831252	12024	184294424	T/C	0.752	0.762	0.695
512071	13963	184296363	C/T	0.616	0.642	0.367
1502761	14300	184296700	A/C	0.596	0.593	0.933
681516	14361	184296761	C/T	0.240	0.189	<b>0.037</b>
619424	16287	184298687	T/G	0.076	0.070	0.704
620722	not mapped		C/T	0.779	0.819	0.100
529055	18635	184301035	A/G	0.601	0.637	0.219
664010	19365	184301765	T/G	0.455	0.394	<b>0.039</b>
678454	not mapped		T/G	0.000	0.004	0.117
2653845	24953	184307353	G/A	0.175	0.168	0.775

dbSNP rs#	Position in Fig 3	Chrom Position	Alleles (A1/A2)	A2 Case AF	A2 Control AF	p-Value
472795	25435	184307835	G/A	0.082	0.077	0.756
502289	not mapped		T/G	0.003	0.000	0.172
507079	26847	184309247	G/A	0.833	0.835	0.937
534333	27492	184309892	T/C	0.496	0.509	0.675
831242	27620	184310020	T/C	0.728	0.776	0.064
536111	27678	184310078	C/T	0.800	0.812	0.632
536213	27714	184310114	G/A	0.271	0.281	0.710
831245	29719	184312119	A/G	0.020	0.012	0.314
639690	30234	184312634	T/C	0.117	0.106	0.577
684174	31909	184314309	T/C	0.304	0.298	0.826
571761	32153	184314553	C/G	0.406	0.425	0.525
1983421	33572	184315972	T/C	0.433	0.425	0.791
2314415	42164	184324564	T/G	0.014	0.050	0.001
2103062	43925	184326325	A/G	0.328	0.361	0.256
6804951	45031	184327431	C/T	no AT	no AT	-
1403452	45655	184328055	T/C	0.025	0.072	0.001
903950	48350	184330750	C/A	0.577	0.594	0.556
2017340	48418	184330818	A/G	0.033	0.054	0.089
2001449	48563	184330963	G/C	0.262	0.205	0.025
3821522	53189	184335589	A/G	0.500	0.480	0.508
1390831	56468	184338868	T/G	0.944	0.923	0.160
1353566	59358	184341758	C/A	0.545	0.533	0.692
1813856	63761	184346161	C/T	0.040	0.041	0.933
2272115	65931	184348331	G/A	0.324	0.370	0.106
3732603	67040	184349440	G/C	0.228	0.209	0.429
940055	69491	184351891	A/C	0.225	0.198	0.272
2314730	83308	184365708	A/G	0.649	0.691	0.135
484315	not mapped		C/G	0.256	0.234	0.404
KIAA0861_3732602	126545	184408945	C/T	no AT	no AT	-
KIAA0861_2293203	137592	184419992	A/T	no AT	no AT	-
7639705	147169	184429569	G/T	no AT	no AT	-

[0285] Figure 16 shows the proximal SNPs in and around the KIAA0861 gene for females. As indicated, some of the SNPs were untyped. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 16 can be determined by consulting Table 23. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0286] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter



8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than  $10^{-8}$  were truncated at that value.

[0287] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link ([http address: www.ncbi.nlm.nih.gov/LocusLink/](http://www.ncbi.nlm.nih.gov/LocusLink/)) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is place at the 3' end of each gene to show the direction of transcription.

#### Additional Genotyping

[0288] A total of five SNPs, including the incident SNP, were genotyped in the discovery cohort. The discovery cohort is described in Example 1. Four of the SNPs are non-synonomous, coding SNPs. Two of the SNPs (rs2001449 and rs6804951) were found to be significantly associated with breast cancer with a p-value of 0.001 and 0.007, respectively. See Table 26.

[0289] The methods used to verify and genotype the five proximal SNPs of Table 26 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 24 and Table 25, respectively.

**Table 24**

dbSNP rs#	Forward PCR primer	Reverse PCR primer
rs7639705	ACGTTGGATGTGTCAGAAAGCAAACCTGGC	ACGTTGGATGTTACAGGCATTGGAGACAGC
rs2293203	ACGTTGGATGCTGCATAATGGTGGCTTTGG	ACGTTGGATGTGTGGGTGTTCACTTTGCAG
rs3732602	ACGTTGGATGCCCTCTTGTGTCAGGAAGTTCT	ACGTTGGATGGAGACAGAGTTGAACTCCCG
rs2001449	ACGTTGGATGAGGAAGAAACTGACGGAAGG	ACGTTGGATGATGTCAAGTGCACCCACATG
rs6804951	ACGTTGGATGAAGATACGAATGGAGCCTGG	ACGTTGGATGGCAATAGGACTCCCTTTACC

**Table 25**

dbSNP rs#	Extend Primer	Term Mix
rs7639705	TGATGCACGTGGAGCAG	CGT
rs2293203	GCCCCTGGAAAAGGCC	CGT
rs3732602	GGAAGATGATGAGACTAAAT	ACG

rs2001449	CACATGCCTGCTCGCCCCC	ACT
rs6804951	TCCCTTTACCTTCATGG	ACG

[0290] Table 26, below, shows the case and control allele frequencies along with the p-values for all of the SNPs genotyped. The disease associated allele of column 4 is in bold and the disease associated amino acid of column 5 is also in bold. The chromosome positions provided correspond to NCBI's Build 33. The amino acid change positions provided in column 5 correspond to KIAA0861 polypeptide sequence of Figure 12.

**Table 26: Genotyping Results**

Rs number	Position in Figure 1	Location within Gene	Alleles (A1/A2)	Amino Acid Change	A2 Case AF	A2 Control AF	p-Value	Odds Ratio
rs7639705	147169	Exon 7	<b>G/T</b>	I276L	0.805	0.811	0.794	1.04
rs2293203	137592	Exon 8	<b>A/T</b>	Q295L	0.990	0.980	0.685	1.25
rs3732602	126545	Exon 11	<b>C/T</b>	S506F	monomorphic			
rs2001449	48563	Intron 19	<b>G/C</b>	-	0.307	0.218	<b>0.001</b>	1.59
rs6804951	45031	Exon 20	<b>C/T</b>	A819T	0.044	0.085	<b>0.007</b>	2.02

#### Example 7

##### NUMA1 Proximal SNPs

[0291] It has been discovered that a polymorphic variation (rs673478) in the NUMA1/FLJ20625/LOC220074 region is associated with the occurrence of breast cancer (see Examples 1 and 2). Subsequently, SNPs proximal to the incident SNP (rs673478) were identified and allelotyped in breast cancer sample sets and control sample sets as described in Examples 1 and 2. Approximately sixty-three allelic variants located within the NUMA1/FLJ20625/LOC220074 region were identified and allelotyped. The polymorphic variants are set forth in Table 27. The chromosome position provided in column four of Table 27 is based on Genome "Build 33" of NCBI's GenBank.

**Table 27**

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
1894003	11	174	71972974	T/C
2390981	11	815	71973615	G/A
1939242	11	3480	71976280	C/T
1894004	11	9715	71982515	T/C
645603	11	14755	71987555	G/A
661290	11	15912	71988712	A/G
679926	11	19834	71992634	A/G

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
567026	11	19850	71992650	G/A
678193	11	20171	71992971	T/G
560777	11	20500	71993300	C/T
676721	11	20536	71993336	C/T
585228	11	23187	71995987	C/G
674319	11	25289	71998089	C/T
675185	11	25470	71998270	T/G
575871	11	28720	72001520	A/G
547208	11	29566	72002366	C/T
2511075	11	30155	72002955	T/C
642573	11	30752	72003552	C/G
671681	11	32710	72005510	C/T
541022	11	32954	72005754	A/G
2511076	11	33725	72006525	G/A
3018308	11	33842	72006642	T/C
671132	11	36345	72009145	G/A
552966	11	38115	72010915	A/C
607446	11	39150	72011950	C/T
3018302	11	40840	72013640	T/G
3018301	11	41969	72014769	A/G
2511114	11	42045	72014845	C/T
548961	11	43785	72016585	G/A
575831	11	44444	72017244	A/G
577435	11	44579	72017379	T/C
495567	11	45386	72018186	C/T
493065	11	46827	72019627	A/G
597513	11	47320	72020120	A/T
598835	11	47625	72020425	T/C
610004	11	47837	72020637	T/C
610041	11	47866	72020666	A/G
673478	11	49002	72021802	T/C
670802	11	49566	72022366	T/G
2511116	11	52058	72024858	C/T
NUMA1_SNP1	11	52249	72025049	A/C
517837	11	52257	72025057	C/T
615000	11	52850	72025650	T/G
482013	11	53860	72026660	C/T
NUMA1_SNP2	11	54052	72026852	T/C
2250866	11	54411	72027211	T/C
2511078	11	55098	72027898	G/A
2508858	11	55303	72028103	C/G
681069	11	59398	72032198	A/G
595062	11	59533	72032333	A/G
542752	11	60542	72033342	A/T
2508856	11	61541	72034341	C/T
832658	11	62309	72035109	G/A
3750908	11	72299	72045099	C/T

dbSNP rs#	Chromosome	Position in Figure 4	Chromosome Position	Allele Variants
3793938	11	73031	72045831	C/T
2276396	11	73803	72046603	G/C
1806778	11	80950	72053750	T/C
4073394	11	82137	72054937	A/G
471547	11	96077	72068877	G/T
606136	11	96470	72069270	A/G
532360	11	98116	72070916	G/T
703781	11	98184	72070984	A/C
476753	11	132952	72105752	A/G

#### Assay for Verifying and Allelotyping SNPs

[0292] The methods used to verify and allelotype the proximal SNPs of Table 27 are the same methods described in Examples 1 and 2 herein. The PCR primers and extend primers used in these assays are provided in Table 28 and Table 29, respectively.

**Table 28**

dbSNP rs#	Forward PCR primer	Reverse PCR primer
744293	ACGTTGGATGTCTGCAGACAGTGGCCAATG	ACGTTGGATGAGGGCCCCAGGATCACAATAG
750789	ACGTTGGATGTTTCATCTGGTAAGTCCCACC	ACGTTGGATGTGAAACAAGAGAGGCCCTTC
1939110	ACGTTGGATGTCTTTAGGTCCAGGATTCCC	ACGTTGGATGTATAGTCAGCATCGTCCCTG
2005192	ACGTTGGATGCCCTCAGAGTTTGGACATAT	ACGTTGGATGTATCCAAAATGCAGACACAG
SNP0000 4859	ACGTTGGATGGTGTATTATCCCAACCCTTCC	ACGTTGGATGGGAGGAAATACAGCCTGTTC
744292	ACGTTGGATGATCCTAGAGGACTGGGAAAG	ACGTTGGATGCTGCTTCTGTTCCCACAATG
754490	ACGTTGGATGAAGGGTGGAGAACTCATGGG	ACGTTGGATGACCCCTATTTTGAAGCAGGC
872619	ACGTTGGATGTTACACCAAGGTGTTACTG	ACGTTGGATGCACAATAATGTGTTCCAGGGC
1807014	ACGTTGGATGCTGGGCAACAAGAGTGAAC	ACGTTGGATGGCCCAAAACCACTGAGATTC
1815753	ACGTTGGATGTAGAGTGAAGACAGAGCTCC	ACGTTGGATGATAAACCCAGGCATTTCGAGC
1892893	ACGTTGGATGTCCTATGAAGATTCTATCTGC	ACGTTGGATGGTCCAGAGTTTTAGACTCAAG
1939111	ACGTTGGATGTCCTTAACCTTATTGGTGGC	ACGTTGGATGGTTGGGTTTCAGTAGAAGAGA
1939112	ACGTTGGATGAGCCACCAATAAGGTTAAGG	ACGTTGGATGTGTCTCTCACTTCCTCAACC
1939113	ACGTTGGATGAGACACACAAGGCAAGGTTTC	ACGTTGGATGCCAGAGAGGAGTCTGTCTAG
1939114	ACGTTGGATGGAAAACATTGGTCCAGGCAG	ACGTTGGATGCAAGAACCCAGGCATCAATG
1939115	ACGTTGGATGGACCACGGAATCCTTTTTTCA	ACGTTGGATGGCTCAAATTCTGTTCTTTAG
1939116	ACGTTGGATGACATAGGTAGTCAGGCACTC	ACGTTGGATGGCAGCTCTTTTTTCTACC
1939117	ACGTTGGATGGGGAACTTTTACATTACAC	ACGTTGGATGGAGAGTTTGCATTGGTGATC
1939118	ACGTTGGATGATGTTGCTGTATGGTCTCTCC	ACGTTGGATGGAAAACATTGCGCTAGGCAC
1954769	ACGTTGGATGTGAGTGACCAAGTTGCTCTG	ACGTTGGATGTCTACCTTCATGATGTCCCC
2000537	ACGTTGGATGGGTCTTTTATGAGGTTTCTCC	ACGTTGGATGGTTAAACTTACAAATCTAGC
2011913	ACGTTGGATGGCTGAGTGTGGATTGCTCTG	ACGTTGGATGAGTAAACCAACACCCAGAAC
2015747	ACGTTGGATGTGAAGCAGGCTTTCCCAATG	ACGTTGGATGGGTAGTGAAGGGTGGAGAAC
2105587	ACGTTGGATGAAGAAATACCAGGCCGGGAG	ACGTTGGATGCTCAAGTATCCTCCCTTCTC
2155081	ACGTTGGATGAGGCAATGCTTCCATTGTTTC	ACGTTGGATGTCATAGCATTTTACCCCTGG

dbSNP rs#	Forward PCR primer	Reverse PCR primer
2186617	ACGTTGGATGGCTACATATGGATCTTGGTC	ACGTTGGATGGACCAGCACTAACTCTAAAC
2508423	ACGTTGGATGCTCCTCTGTAAACCAGGAC	ACGTTGGATGAGAACTCTCCTAAGCACAC
2511880	ACGTTGGATGGTTCCCTGATGGAAAATGCC	ACGTTGGATGCCAGAATGCCTTATCCACAG
2511881	ACGTTGGATGTGACTCTGCTGTGAGATTGG	ACGTTGGATGACATCGGTTTCACCTCCAAC
2512990	ACGTTGGATGAGCCAGCAGAGAAAACAGTC	ACGTTGGATGGCCACTTACTACCTGTTGTC
2555537	ACGTTGGATGGGACATAACCATAGGCCATC	ACGTTGGATGCATTGACAGCTGTATTGCAC
3016250	ACGTTGGATGTTTTTGAGACGGAGTCTCGC	ACGTTGGATGAGGCAGGAGAATGGCGTGAA
3016251	ACGTTGGATGAGCTTGCACTGAGCCGAGAT	ACGTTGGATGTTTTTGAGACGGAGTCTCGC
3016252	ACGTTGGATGTGGTGAAGAGAAGTCAAAGC	ACGTTGGATGAGGCTGAATGATTCCCCTTC
3781614	ACGTTGGATGTGGTCAGTCAGTTAGCCAGG	ACGTTGGATGCCCTAATGATGGTAGACTGC
3809048	ACGTTGGATGACCACCAAGATAACGACCGC	ACGTTGGATGAGCCACCTCCTTGTCAGTG
4128368	ACGTTGGATGGGACAATATTTAGTTATGCAC	ACGTTGGATGTTCAAGGTCATCCCCTTATC

Table 29

dbSNP rs#	Extend Primer	Term Mix
744293	GATGGCCCAGTTCCTGCGC	ACG
750789	AGAGGCCCTTCCAGGGCT	ACT
1939110	CGTCCCTGACCTGGACTTA	ACG
2005192	AATGCAGACACAGTTCTGGG	CGT
SNP00004859	CTGAAAAATAGCTAGTTC	ACG
744292	ACTCACCTCTACCCATAAGG	ACT
754490	TTGAAGCAGGCTTTCCCA	ACT
872619	TGTGTTCAAGGCTTTCTCAT	ACT
1807014	GTGTTTTTTTTTCCCCC	ACG
1815753	CAGGCATTCGAGCCAGCAAT	ACT
1892893	ATGTTTTATTCTTTCACAAAAGT	ACT
1939111	GGAGGAGGCAGTAAGGAA	ACT
1939112	CTTCCAACTTTTTCTCTTG	ACT
1939113	GTCTAGTCCTCCAAGCC	ACG
1939114	ATCAATGGGGTGGTGCA	ACT
1939115	TCTGTTCTTTAGAAGGCT	CGT
1939116	TGTACCAATATGACAATTTAACC	ACT
1939117	CCTGACACATAGTTCATGCTC	ACT
1939118	GCTAGGCACAAAATTAAAGAGAT	ACT
1954769	TCCCCGCCTTTCCTCC	CGT
2000537	ACAAATCTAGCACCGAAGG	ACT
2011913	ATATAAGCAATTCACAAGTAATGT	ACT
2015747	AAGGGTGGAGAACTCATGG	ACT
2105587	TATCCTCCCTTCTCAGCAAG	ACT
2155081	CATTTTACCCCTGGATTATA	ACT
2186617	CTCAACCTCAACTCAACT	CGT
2508423	TCTCCTAAGCACACTATGTATAT	ACG
2511880	AGGATATTAGTCATGCTGGG	ACT
2511881	CACCTCCAACACGGTCCCC	CGT

2512990	GTTGTCTTCCCAACTCC	ACT
2555537	ACTGTGGACATTGGTGT	ACT
3016250	GGCGTGAACCCGGGAGG	ACG
3016251	CTGTCGCCCAGGCCGGA	ACT
3016252	GATTCCTTCTTCTAAA	ACT
3781614	TAGACTGCAGAGTAGCA	ACT
3809048	TGGGCCTACTTCCCTGA	ACT
4128368	TTTTCATCACATAGCTCATCT	CGT

### Genetic Analysis of Allelotyping Results

[0293] Allelotyping results are shown for cases and controls in Table 30. The allele frequency for the A2 allele is noted in the fifth and sixth columns for breast cancer pools and control pools, respectively, where “AF” is allele frequency. The allele frequency for the A1 allele can be easily calculated by subtracting the A2 allele frequency from 1 (A1 AF = 1-A2 AF). For example, the SNP rs1894003 has the following case and control allele frequencies: case A1 (T) = 0.192; case A2 (C) = 0.808; control A1 (T) = 0.115; and control A2 (C) = 0.885, where the nucleotide is provided in paranthesis. SNPs with blank allele frequencies were untyped.

**Table 30**

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
1894003	174	71972974	T/C	0.808	0.885	0.00061
2390981	815	71973615	G/A	0.013	0.002	0.02306
1939242	3480	71976280	C/T	0.902	0.943	0.01186
1894004	9715	71982515	T/C	0.020	0.009	0.12637
645603	14755	71987555	G/A	0.029	0.021	0.37479
661290	15912	71988712	A/G	0.813	0.833	0.39013
679926	19834	71992634	A/G	0.077	0.039	0.00741
567026	19850	71992650	G/A	0.059	0.038	0.09767
678193	20171	71992971	T/G	0.868	0.920	0.00597
560777	20500	71993300	C/T	0.070	0.041	0.03071
676721	20536	71993336	C/T	0.901	0.947	0.00419
585228	23187	71995987	C/G	0.842	0.914	0.00043
674319	25289	71998089	C/T	0.027	0.027	0.96556
675185	25470	71998270	T/G	0.763	0.853	0.00031
575871	28720	72001520	A/G	0.924	0.932	0.61199
547208	29566	72002366	C/T	0.042	0.023	0.07555
2511075	30155	72002955	T/C	0.894	0.944	0.00256
642573	30752	72003552	C/G	0.047	0.022	0.02382
671681	32710	72005510	C/T	0.072	0.043	0.03643
541022	32954	72005754	A/G	0.070	0.040	0.02829
2511076	33725	72006525	G/A	0.223	0.256	0.20380
3018308	33842	72006642	T/C	0.442	0.439	0.92279
671132	36345	72009145	G/A	0.970	0.971	0.96469
552966	38115	72010915	A/C	0.845	0.903	0.00393
607446	39150	72011950	C/T	0.861	0.918	0.00279
3018302	40840	72013640	T/G	0.767	0.827	0.01378
3018301	41969	72014769	A/G	0.734	0.837	0.00011

dbSNP rs#	Position in Figure 4	Chromosome Position	A1/A2 Allele	A2 Case AF	A2 Control AF	p-Value
2511114	42045	72014845	C/T	0.080	0.036	0.00222
548961	43785	72016585	G/A	0.852	0.905	0.00833
575831	44444	72017244	A/G	0.946	0.961	0.22995
577435	44579	72017379	T/C	0.013	0.007	0.34863
495567	45386	72018186	C/T	0.891	0.951	0.00045
493065	46827	72019627	A/G	0.823	0.904	0.00022
597513	47320	72020120	A/T	0.890	0.936	0.00667
598835	47625	72020425	T/C	0.074	0.038	0.00994
610004	47837	72020637	T/C	0.088	0.041	0.00209
610041	47866	72020666	A/G	0.872	0.933	0.00102
673478	49002	72021802	T/C	0.173	0.094	0.00026
670802	49566	72022366	T/G	0.876	0.920	0.01646
2511116	52058	72024858	C/T	0.898	0.945	0.00437
NUMA1_SNP1	52249	72025049	A/C	0.901	0.924	0.17421
517837	52257	72025057	C/T	0.095	0.061	0.03504
615000	52850	72025650	T/G	0.812	0.916	0.00001
482013	53860	72026660	C/T	0.884	0.924	0.02391
NUMA1_SNP2	54052	72026852	T/C	0.066	0.034	0.01392
2250866	54411	72027211	T/C	0.855	0.918	0.00132
2511078	55098	72027898	G/A	0.299	0.295	0.86946
2508858	55303	72028103	C/G	0.898	0.944	0.00509
681069	59398	72032198	A/G	0.835	0.878	0.04069
595062	59533	72032333	A/G	0.925	0.942	0.25198
542752	60542	72033342	A/T	0.853	0.915	0.00192
2508856	61541	72034341	C/T	0.074	0.060	0.33745
832658	62309	72035109	G/A	0.047	0.023	0.02994
3750908	72299	72045099	C/T	0.912	0.944	0.04342
3793938	73031	72045831	C/T	0.084	0.045	0.00763
2276396	73803	72046603	G/C	0.892	0.937	0.00799
1806778	80950	-72053750	T/C	0.041	0.034	0.50886
4073394	82137	72054937	A/G	0.547	0.579	0.28705
471547	96077	72068877	G/T	0.490	0.522	0.28304
606136	96470	72069270	A/G	0.444	0.468	0.43474
532360	98116	72070916	G/T	0.043	0.021	0.03475
703781	98184	72070984	A/C	0.078	0.080	0.89053
476753	132952	72105752	A/G	0.922	0.936	0.39563

[0294] Figure 17 shows the proximal SNPs in and around the *NUMA1* region for females. The position of each SNP on the chromosome is presented on the x-axis. The y-axis gives the negative logarithm (base 10) of the p-value comparing the estimated allele in the case group to that of the control group. The minor allele frequency of the control group for each SNP designated by an X or other symbol on the graphs in Figure 17 can be determined by consulting Table 30. By proceeding down the Table from top to bottom and across the graphs from left to right the allele frequency associated with each symbol shown can be determined.

[0295] To aid the interpretation, multiple lines have been added to the graph. The broken horizontal lines are drawn at two common significance levels, 0.05 and 0.01. The vertical broken lines are drawn every 20kb to assist in the interpretation of distances between SNPs. Two other lines are drawn to expose linear trends in the association of SNPs to the disease. The light gray line (or generally bottom-

most curve) is a nonlinear smoother through the data points on the graph using a local polynomial regression method (W.S. Cleveland, E. Grosse and W.M. Shyu (1992) Local regression models. Chapter 8 of Statistical Models in S eds J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.). The black line (or generally top-most curve, *e.g.*, see peak in left-most graph just to the left of position 92150000) provides a local test for excess statistical significance to identify regions of association. This was created by use of a 10kb sliding window with 1kb step sizes. Within each window, a chi-square goodness of fit test was applied to compare the proportion of SNPs that were significant at a test wise level of 0.01, to the proportion that would be expected by chance alone (0.05 for the methods used here). Resulting p-values that were less than  $10^{-8}$  were truncated at that value.

[0296] Finally, the gene or genes present in the loci region of the proximal SNPs as annotated by Locus Link ([http address: www.ncbi.nlm.nih.gov/LocusLink/](http://www.ncbi.nlm.nih.gov/LocusLink/)) are provided on the graph. The exons and introns of the genes in the covered region are plotted below each graph at the appropriate chromosomal positions. The gene boundary is indicated by the broken horizontal line. The exon positions are shown as thick, unbroken bars. An arrow is placed at the 3' end of each gene to show the direction of transcription.

#### Example 8

##### Meta Analysis of Incident SNPs

[0297] Meta-analysis was performed of five of the incident SNPs disclosed in Table 3 (ICAM region (ICAM\_SNP), MAPK10 (rs1541998), KIAA0861 (rs2001449), NUMA1 region (rs673478) and GALE region (rs4237)) based on genotype results provided in Table 6B. Figures 18-21 depict odds ratios for the discovery samples and replication samples (see Example 3) individually, and the combined meta analysis odds ratio for the named SNP. The boxes are centered over the odds ratio for each sample, with the size of the box correlated to the contribution of each sample to the combined meta analysis odds ratio. The lines extending from each box are the 95% confidence interval values. The diamond is centered over the combined meta analysis odds ratio with the ends of the diamond depicting the 95% confidence interval values. The meta-analysis further illustrates the strong association each of the incident SNPs has with breast cancer across multiple case and control samples.

[0298] The subjects available for discovery from Germany included 272 cases and 276 controls. The subjects available for replication from Australia included 190 breast cancer cases and 190 controls. Meta analyses, combining the results of the German discovery sample and the Australian replication sample, were carried out using a random effects (DerSimonian-Laird) procedure.



Example 9

Description of development of predictive breast cancer models

[0299] The five SNPs reported in Example 3 were identified as being significantly associated with breast cancer according to the replication analysis discussed therein. These five SNPs are a subset of the panel of SNPs associated with breast cancer in the German chort referenced in Example 1 and reported in provisional patent application no. 60/429,136 filed November 25, 2002 and provisional patent application no. 60/490,234 filed July 24, 2003, having attorney docket number 524593004100 and 524593004101, respectively.

[0300] The clinical importance of these SNPs was estimated by combining them into a single logistic regression model. The coefficients of the model were used to estimate penetrance, relative risk and odds ratio values for estimating a subject's risk of having or developing breast cancer according to the subject's genotype. Penetrance is a probability that an individual has or will have breast cancer given their genotype (e.g., a value of 0.01 in the tables is equal to a 1% chance of having or developing breast cancer). The relative risk of breast cancer is based upon penetrance values, and is expressed in two forms. One form, noted as RR in the tables below, is expressed as a risk with respect to the lowest risk group (e.g., the most protected group being the 00000 genotype listed in Table 33). The other form is expressed as a risk with respect to a population average risk of breast cancer, which is noted as RR(Pop) in Table 35 below. Both of these expressions of relative risk are useful to a clinician for assessing risk of breast cancer in an individual and targeting appropriate detection, prevention and/or treatment regimens to the subject. Both expressions of relative risk also are useful to an insurance company to assess population risks of breast cancer (e.g., for developing actuarial tables), where individual genotypes often are provided to the company on an anonymous basis. Odds ratios are the odds one group has or will develop breast cancer with respect to another group, the other group often being the most protective group or the group having a population average risk of breast cancer. Relative risk often is a more reliable assessment of risk in comparison to an odds ratio when the disease or condition at issue is more prevalent.

[0301] To fit the single logistic model, all cases and controls from the German and Australian samples were used (see Examples 1 and 3, respectively). Controls were coded as 0 and cases were coded as 1. Based on the genotype penetrance estimates of each SNP (Table 31), GP01.025495354 (rs4237), GP03.197942797 (rs2001449), GP11.079035103 (rs673478) were modeled as additive by coding the genotypes 0, 1, or 2 for the low risk homozygote, the heterozygote, or high risk homozygote, respectively. The SNP FCH.0994 (ICAM\_SNP1) was modeled as recessive coding the genotypes 0, 0, or 2 for the low risk homozygote, heterozygote, or high risk homozygote, respectively. The SNP GP04.091348915 (rs1541998) was modeled as dominant coding the genotypes 0, 2, or 2 for the low risk

homozygote, the heterozygote, or high risk homozygote, respectively. Table 31 summarizes this analysis.

**Table 31**

SNP: Genotype	N	Case (N=254)	Control (N=268)	P(D G) (%)	P-value
ICAM_SNP1:	497	45%(103)	32% (85)	4.140	0.006210
CC		42%(98)	47% (126)	2.700	
CT		13% (30)	21% (13)	1.910	
TT					
rs4237: AA	494	34% (79)	29% (75)	3.550	0.186000
AG		49% (113)	48% (126)	3.040	
GG		17% (40)	23% (61)	2.240	
rs2001449:	508	46% (112)	60% (158)	2.280	0.002930
GG		48% (117)	36% (94)	3.940	
GC		7% (17)	4% (10)	5.300	
CC					
rs673478: TT	509	84% (206)	91% (240)	2.800	0.040700
TC		14% (35)	9% (25)	4.490	
CC		1% (3)	0% (0)	100.00	
rs1541998:	493	5% (12)	4% (10)	3.710	0.012100
CC		36% (87)	24% (61)	4.370	
CT		59% (143)	72% (180)	2.490	
TT					

[0302] Based on this coding, there are a total of 108 unique genotype codes from the 243 unique five SNP genotypes. The relationship between the five SNP genotypes and the case-control status was fit using logistic regression. Many models were fit and compared including the five SNPs and all possible interaction among SNPs and study center. Only statistically significant terms from this complete model were included in the final model, shown in Table 32.

**Table 32**

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.34446	0.25972	-5.177	2.26e-07
FCH.0994	0.77607	0.19835	3.913	9.13e-05
4237	0.54525	0.17666	3.086	0.002025
2001449	0.60383	0.28487	2.120	0.034033
1541998	0.22051	0.07849	2.809	0.004963
673478	0.59961	0.21737	2.758	0.005807
FCH.0994c: 4237	-0.52636	0.14516	-3.626	0.000288
FCH.0994c: 2001449	-0.35613	0.24503	-1.453	0.146113
4237c: 2001449	-0.15685	0.20191	-0.777	0.437257
FCH.0994c: 4237c2001449	0.41305	0.18391	2.246	0.024705

Null deviance: 1136.7 on 820 degrees of freedom

Residual deviance: 1069.6 on 811 degrees of freedom

AIC: 1089.6

[0303] The penetrance was calculated for each of the 108 unique genotype codes using this model and an assumed disease prevalence of 0.03 (prev), the cumulative incidence for the age range of the sample in question. This was calculated from the logistic model as follows:

$$\text{penetrance} = \exp(\hat{y} + \text{adj}) / (1 + \exp(\hat{y} + \text{adj}))$$

where

$$\hat{y} = 1 / (1 + \exp(-1.344 + 0.776*A + 0.545*B + 0.604*C + 0.221*D + 0.600*E - 0.526*A*B - 0.356*A*C - 0.157*B*C + 0.413*A*B*C))$$

and

$$\text{adj} = \ln(\text{prev} / (1 - \text{prev}) * \text{freq}(\text{case}) / (1 - \text{freq}(\text{case}))).$$

Here A, B, C, D, and E refer to the genotype codes for the SNPs FCH.0994, 4237, 2001449, 1541998, and 673478, respectively.

[0304] Table 33 summarizes statistics of interest for each genotype code. “Geno” shows each genotype code with the five integer codes formatted as an integer string. “N Case” and “N Control” is the number of cases and controls with the specified code, respectively. “Frequency” is the expected percent of individuals in the population having that code calculated as the average of the case and control frequencies weighted by the probability of disease in this sample (0.03). “OR” is the odds ratio comparing the odds of the specified code to the odds of the most protective code (00000) using the parameter estimates from the logistic regression model. “OR (Frq)” is an odds ratio estimated using the frequency of cases and control with the specified genotype code and the most protective code. “RR” is the relative risk comparing the probability of disease of the specified code to the probability of disease of the most protective code. “Penetrance” is the probability of disease given the genotype code, followed by “Lower” and “Upper” which give the 95% confidence interval for the penetrance. As can be seen by the ratios for OR and RR, the 00000 genotype was the most protective against breast cancer occurrence.

**Table 33**

Geno	N Case	N Control	Frequency	OR	OR (Frq)	RR	Penetrance	Confidence Interval	
								Lower	Upper
00000	6	26	5.94%	1.00	1.00	1.00	0.010	0.007	0.014
00001	0	3	0.68%	1.75	0.00	1.74	0.017	0.011	0.029
00002	0	0	0.00%	3.08		3.01	0.030	0.013	0.069
00020	3	9	2.06%	1.61	1.44	1.60	0.016	0.011	0.023
00021	0	3	0.68%	2.83	0.00	2.78	0.028	0.017	0.047
00022	0	0	0.00%	4.97		4.78	0.048	0.021	0.108
00100	9	20	4.60%	1.67	1.95	1.66	0.017	0.012	0.023
00101	2	1	0.24%	2.93	8.67	2.87	0.029	0.018	0.047

Geno	N Case	N Control	Frequency	OR	OR (Frq)	RR	Penetrance	Confidence Interval	
								Lower	Upper
00102	0	0	0.00%	5.13		4.93	0.050	0.022	0.110
00120	7	6	1.41%	2.69	5.06	2.65	0.027	0.018	0.038
00121	0	0	0.00%	4.73		4.56	0.046	0.028	0.075
00122	0	0	0.00%	8.29		7.72	0.078	0.034	0.168
00200	1	4	0.91%	2.78	1.08	2.74	0.027	0.018	0.042
00201	0	0	0.00%	4.88		4.70	0.047	0.027	0.082
00202	0	0	0.00%	8.57		7.96	0.080	0.034	0.178
00220	1	1	0.23%	4.50	4.33	4.34	0.044	0.027	0.070
00221	1	0	0.01%	7.89		7.38	0.074	0.041	0.129
00222	0	0	0.00%	13.83		12.25	0.123	0.052	0.263
01000	24	47	10.84%	1.26	2.21	1.26	0.013	0.010	0.016
01001	3	1	0.25%	2.21	13.00	2.18	0.022	0.014	0.034
01002	0	0	0.00%	3.87		3.77	0.038	0.017	0.083
01020	18	22	5.12%	2.03	3.55	2.01	0.020	0.015	0.027
01021	4	4	0.94%	3.57	4.33	3.48	0.035	0.022	0.055
01022	0	0	0.00%	6.26		5.94	0.060	0.027	0.129
01100	21	33	7.64%	2.10	2.76	2.08	0.021	0.017	0.026
01101	2	4	0.92%	3.69	2.17	3.59	0.036	0.024	0.055
01102	0	0	0.00%	6.47		6.13	0.062	0.028	0.130
01120	15	6	1.47%	3.39	10.83	3.31	0.033	0.025	0.045
01121	0	0	0.00%	5.95		5.67	0.057	0.036	0.089
01122	0	0	0.00%	10.44		9.54	0.096	0.044	0.198
01200	5	4	0.94%	3.51	5.42	3.42	0.034	0.023	0.050
01201	0	1	0.23%	6.15	0.00	5.85	0.059	0.035	0.097
01202	0	0	0.00%	10.79		9.82	0.099	0.044	0.209
01220	1	0	0.01%	5.66		5.41	0.054	0.035	0.083
01221	0	0	0.00%	9.93		9.12	0.092	0.054	0.152
01222	0	0	0.00%	17.42		14.95	0.150	0.067	0.304
02000	22	39	9.01%	1.59	2.44	1.58	0.016	0.012	0.021
02001	2	1	0.24%	2.78	8.67	2.73	0.027	0.017	0.043
02002	1	0	0.01%	4.88		4.70	0.047	0.021	0.103
02020	16	10	2.39%	2.56	6.93	2.52	0.025	0.018	0.035
02021	2	2	0.47%	4.49	4.33	4.34	0.044	0.027	0.070
02022	2	0	0.02%	7.88		7.37	0.074	0.033	0.158
02100	21	18	4.24%	2.65	5.06	2.60	0.026	0.020	0.035
02101	5	3	0.72%	4.64	7.22	4.48	0.045	0.029	0.070
02102	0	0	0.00%	8.14		7.60	0.076	0.035	0.160
02120	11	8	1.90%	4.28	5.96	4.14	0.042	0.030	0.058
02121	1	0	0.01%	7.50		7.04	0.071	0.044	0.112
02122	0	0	0.00%	13.15		11.72	0.118	0.054	0.239
02200	4	4	0.94%	4.42	4.33	4.27	0.043	0.028	0.065
02201	3	1	0.25%	7.75	13.00	7.26	0.073	0.043	0.121
02202	0	0	0.00%	13.59		12.06	0.121	0.053	0.252
02220	2	1	0.24%	7.13	8.67	6.72	0.068	0.043	0.106
02221	0	0	0.00%	12.51		11.21	0.113	0.065	0.189
02222	0	0	0.00%	21.94		18.13	0.182	0.082	0.358

Geno	N Case	N Control	Frequency	OR	OR (Frq)	RR	Penetrance	Confidence Interval	
								Lower	Upper
20000	9	6	1.43%	1.58	6.50	1.57	0.016	0.011	0.023
20001	0	0	0.00%	2.76		2.72	0.027	0.016	0.045
20002	0	0	0.00%	4.85		4.67	0.047	0.020	0.105
20020	8	4	0.97%	2.54	8.67	2.51	0.025	0.017	0.037
20021	0	0	0.00%	4.46		4.31	0.043	0.026	0.072
20022	0	0	0.00%	7.83		7.33	0.074	0.032	0.161
20100	5	6	1.40%	2.63	3.61	2.59	0.026	0.018	0.037
20101	4	1	0.26%	4.61	17.33	4.45	0.045	0.027	0.072
20102	0	0	0.00%	8.09		7.55	0.076	0.033	0.163
20120	4	1	0.26%	4.25	17.33	4.11	0.041	0.028	0.060
20121	1	0	0.01%	7.45		6.99	0.070	0.042	0.115
20122	0	0	0.00%	13.06		11.65	0.117	0.052	0.242
20200	0	1	0.23%	4.39	0.00	4.24	0.043	0.027	0.066
20201	1	0	0.01%	7.70		7.21	0.072	0.041	0.124
20202	0	0	0.00%	13.50		11.99	0.121	0.052	0.255
20220	0	0	0.00%	7.09		6.68	0.067	0.041	0.108
20221	0	0	0.00%	12.43		11.15	0.112	0.063	0.192
20222	0	0	0.00%	21.80		18.03	0.181	0.080	0.361
21000	22	25	5.83%	1.99	3.81	1.97	0.020	0.015	0.026
21001	3	4	0.93%	3.48	3.25	3.40	0.034	0.022	0.053
21002	1	0	0.01%	6.11		5.81	0.058	0.026	0.125
21020	11	14	3.26%	3.21	3.40	3.14	0.032	0.023	0.043
21021	1	2	0.46%	5.62	2.17	5.37	0.054	0.034	0.085
21022	0	0	0.00%	9.86		9.05	0.091	0.041	0.190
21100	26	24	5.64%	3.31	4.69	3.24	0.033	0.025	0.042
21101	1	2	0.46%	5.81	2.17	5.54	0.056	0.036	0.085
21102	1	0	0.01%	10.19		9.33	0.094	0.043	0.191
21120	16	6	1.48%	5.35	11.56	5.12	0.051	0.037	0.071
21121	4	0	0.03%	9.38		8.65	0.087	0.055	0.135
21122	0	0	0.00%	16.45		14.24	0.143	0.067	0.281
21200	3	1	0.25%	5.53	13.00	5.29	0.053	0.036	0.078
21201	3	0	0.02%	9.69		8.92	0.090	0.054	0.146
21202	0	0	0.00%	17.00		14.65	0.147	0.067	0.295
21220	2	2	0.47%	8.93	4.33	8.27	0.083	0.053	0.127
21221	1	0	0.01%	15.65		13.65	0.137	0.081	0.223
21222	0	0	0.00%	27.46		21.69	0.218	0.101	0.409
22000	13	23	5.31%	2.50	2.45	2.46	0.025	0.018	0.034
22001	4	1	0.26%	4.39	17.33	4.24	0.043	0.027	0.068
22002	0	1	0.23%	7.69	0.00	7.21	0.072	0.032	0.154
22020	3	10	2.29%	4.04	1.30	3.92	0.039	0.027	0.056
22021	1	0	0.01%	7.08		6.67	0.067	0.041	0.107
22022	0	0	0.00%	12.42		11.14	0.112	0.051	0.230
22100	15	5	1.25%	4.17	13.00	4.04	0.041	0.030	0.055
22101	1	0	0.01%	7.32		6.88	0.069	0.044	0.107
22102	0	0	0.00%	12.83		11.47	0.115	0.053	0.232
22120	3	5	1.16%	6.74	2.60	6.37	0.064	0.045	0.091

Gen	N Case	N Control	Frequency	OR	OR (Frq)	RR	Penetrance	Confidence Interval	
								Lower	Upper
22121	3	1	0.25%	11.82	13.00	10.66	0.107	0.066	0.168
22122	0	0	0.00%	20.72		17.30	0.174	0.081	0.333
22200	4	0	0.03%	6.96		6.57	0.066	0.043	0.100
22201	0	0	0.00%	12.21		10.97	0.110	0.065	0.181
22202	0	0	0.00%	21.42		17.77	0.179	0.081	0.348
22220	4	1	0.26%	11.24	17.33	10.19	0.102	0.064	0.160
22221	0	0	0.00%	19.72		16.60	0.167	0.097	0.271
22222	0	0	0.00%	34.58		25.86	0.260	0.122	0.470

[0305] To simplify the interpretation of genotype risk, the 243 unique genotypes were divided into five risk classes on the basis of each estimated penetrance. The levels selected for risk class definitions and the resulting assignment of genotypes into five risk classes is shown in Table 34. The frequency percent of each genotype combination is given in parentheses.

**Table 34**

Class 1 (0, 0.013]	Class 2 (0.013, 0.025]	Class 3 (0.025, 0.042]	Class 4 (0.042, 0.1]	Class 5 (0.1, 1)
00000 ( 5.94)	00001 ( 0.68)	00022 ( 0.00)	00102 ( 0.00)	00222 ( 0.00)
00020 ( 2.06)	00002 ( 0.00)	00121 ( 0.00)	00122 ( 0.00)	01222 ( 0.00)
01000 (10.84)	00021 ( 0.68)	00220 ( 0.23)	00201 ( 0.00)	02022 ( 0.02)
22000 ( 5.31)	00100 ( 4.60)	01002 ( 0.00)	00202 ( 0.00)	02122 ( 0.00)
	00101 ( 0.24)	01021 ( 0.94)	00221 ( 0.01)	02202 ( 0.00)
	00120 ( 1.41)	01101 ( 0.92)	01022 ( 0.00)	02221 ( 0.00)
	00200 ( 0.91)	01120 ( 1.47)	01102 ( 0.00)	02222 ( 0.00)
	01001 ( 0.25)	01200 ( 0.94)	01121 ( 0.00)	20002 ( 0.00)
	01020 ( 5.12)	02001 ( 0.24)	01122 ( 0.00)	20022 ( 0.00)
	01100 ( 7.64)	02020 ( 2.39)	01201 ( 0.23)	20122 ( 0.00)
	02000 ( 9.01)	02100 ( 4.24)	01202 ( 0.00)	20222 ( 0.00)
	21000 ( 5.83)	02200 ( 0.94)	01220 ( 0.01)	21102 ( 0.01)
	22001 ( 0.26)	20000 ( 1.43)	01221 ( 0.00)	21122 ( 0.00)
	22020 ( 2.29)	20100 ( 1.40)	02002 ( 0.01)	21201 ( 0.02)
		20200 ( 0.23)	02021 ( 0.47)	21202 ( 0.00)
		20220 ( 0.00)	02101 ( 0.72)	21221 ( 0.01)
		21001 ( 0.93)	02102 ( 0.00)	21222 ( 0.00)
		21020 ( 3.26)	02120 ( 1.90)	22102 ( 0.00)
		21100 ( 5.64)	02121 ( 0.01)	22121 ( 0.25)
		22002 ( 0.23)	02201 ( 0.25)	22122 ( 0.00)
		22021 ( 0.01)	02220 ( 0.24)	22200 ( 0.03)
		22100 ( 1.25)	20001 ( 0.00)	22201 ( 0.00)
			20020 ( 0.97)	22202 ( 0.00)
			20021 ( 0.00)	22220 ( 0.26)
			20101 ( 0.26)	22221 ( 0.00)

Class 1 (0, 0.013]	Class 2 (0.013, 0.025]	Class 3 (0.025, 0.042]	Class 4 (0.042, 0.1]	Class 5 (0.1, 1)
			20102 ( 0.00)	22222 ( 0.00)
			20120 ( 0.26)	
			20121 ( 0.01)	
			20201 ( 0.01)	
			20202 ( 0.00)	
			20221 ( 0.00)	
			21002 ( 0.01)	
			21021 ( 0.46)	
			21022 ( 0.00)	
			21101 ( 0.46)	
			21120 ( 1.48)	
			21121 ( 0.03)	
			21200 ( 0.25)	
			21220 ( 0.47)	
			22022 ( 0.00)	
			22101 ( 0.01)	
			22120 ( 1.16)	

[0306] With this classification, each genotype was recoded as belonging to their respective class and a logistic regression model was fit with the genotype risk class as a categorical variable. Key summary statistics are summarized in Table 35. Each group is described by the number of cases, number of controls, the estimated risk class population frequency, the odds ratio comparing the odds of the given risk class compared to the odds of the lowest risk class, the penetrance, the relative risk (risk class penetrance divided by most protective risk class penetrance), and the population relative risk (risk class penetrance divided by the disease prevalence: 0.03).

**Table 35**

Risk Class	N Case	N Control	Frequency (%)	OR	Penetrance	RR	RR (Pop)
G1	46	105	24.2	1.0	0.012	1.0	0.41
G2	112	168	38.9	1.5	0.019	1.5	0.62
G3	140	113	26.7	2.8	0.034	2.8	1.13
G4	77	40	9.7	4.4	0.052	4.2	1.73
G5	18	2	0.06	20.5	0.204	16.6	6.79

### Example 10

#### Inhibition of ICAM Gene Expression by Transfection of Specific siRNAs

[0307] RNAi-based gene inhibition was selected as a rapid way to inhibit expression of ICAM1 in cultured cells. siRNA reagents were selectively designed to target the ICAM1 gene. Algorithms useful for designing siRNA molecules specific for ICAM1 gene are disclosed at the http address [www.dharmacon.com](http://www.dharmacon.com). siRNA molecules up to 21 nucleotides in length were utilized.

[0308] Table 31 summarizes the features of the duplexes that were used in the assays to target ICAM1. A non-homologous siRNA reagent (siGL2 control) was used as a negative control, and a non-homologous siRNA reagent (siRNA\_RAD21\_1175 control) shown to inhibit the expression of RAD21 and subsequently inhibit cell proliferation was used as a positive control in all of the assays described herein.

**Table 36**

siRNA	siRNA Target	Sequence Specificity	SEQ ID NO:
ICAM1_293	ICAM1	ACAACCGGAAGGUGUAUGA	
ICAM1_335	ICAM1	GCCAACCAAUGUGCUAUUC	
ICAM1_604	ICAM1	GAUCACCAUGGAGCCAAUU	
ICAM1_1409	ICAM1	CUGUCACUCGAGAUUCUUGA	
siRNA_RAD21_1175 positive control	RAD21	GAGUUGGAUAGCAAGACAA	
siGL2 negative control	GL2	CGUACGCGGAUACUUCGA	

[0309] The siRNAs were transfected in cell lines MCF-7 and T-47D using Lipofectamine™ 2000 reagent from Invitrogen, Corp. 2.5 µg or 5.0 µg of siRNA was mixed with 6.25 µl or 12.5 µl lipofectamine, respectively, and the mixture was added to cells grown in 6-well plates. Their inhibitory effects on ICAM1 gene expression were confirmed by precision expression analysis by MassARRAY (quantitativeRT-PCR hME), which was performed on RNA prepared from the transfected cells. See Chunming & Cantor, *PNAS* 100(6):3059-3064 (2003). Cell viability was measured at 1, 2, 4 and 6 days post-transfection. Absorbance values were normalized relative to Day 1. RNA was extracted with Trizole reagent as recommended by the manufacturer (Invitrogen, Corp.) followed by cDNA synthesis using SuperScript™ reverse transcriptase.

[0310] A cocktail of siRNA molecules described in Table 28 (that target ICAM1) strongly inhibited proliferation of breast cancer cell line (MCF-7), as shown in in Figure 22. These effects are consistent in all six experiments performed. Each data point is an average of 3 wells of a 96-well plate normalized to values obtained from day 1 post transfection. The specificity of the active siRNAs, was confirmed with



a negative, non-homologous control siRNA (siGL2), and a positive control, siRNA\_RAD21\_1175, that targets a known cancer-associated gene, RAD21.

#### Example 11

##### In Vitro Production of Target Polypeptides

[0311] cDNA is cloned into a pIVEX 2.3-MCS vector (Roche Biochem) using a directional cloning method. A cDNA insert is prepared using PCR with forward and reverse primers having 5' restriction site tags (in frame) and 5-6 additional nucleotides in addition to 3' gene-specific portions, the latter of which is typically about twenty to about twenty-five base pairs in length. A Sal I restriction site is introduced by the forward primer and a Sma I restriction site is introduced by the reverse primer. The ends of PCR products are cut with the corresponding restriction enzymes (*i.e.*, Sal I and Sma I) and the products are gel-purified. The pIVEX 2.3-MCS vector is linearized using the same restriction enzymes, and the fragment with the correct sized fragment is isolated by gel-purification. Purified PCR product is ligated into the linearized pIVEX 2.3-MCS vector and *E. coli* cells transformed for plasmid amplification. The newly constructed expression vector is verified by restriction mapping and used for protein production.

[0312] *E. coli* lysate is reconstituted with 0.25 ml of Reconstitution Buffer, the Reaction Mix is reconstituted with 0.8 ml of Reconstitution Buffer; the Feeding Mix is reconstituted with 10.5 ml of Reconstitution Buffer; and the Energy Mix is reconstituted with 0.6 ml of Reconstitution Buffer. 0.5 ml of the Energy Mix was added to the Feeding Mix to obtain the Feeding Solution. 0.75 ml of Reaction Mix, 50  $\mu$ l of Energy Mix, and 10  $\mu$ g of the template DNA is added to the *E. coli* lysate.

[0313] Using the reaction device (Roche Biochem), 1 ml of the Reaction Solution is loaded into the reaction compartment. The reaction device is turned upside-down and 10 ml of the Feeding Solution is loaded into the feeding compartment. All lids are closed and the reaction device is loaded into the RTS500 instrument. The instrument is run at 30°C for 24 hours with a stir bar speed of 150 rpm. The pIVEX 2.3 MCS vector includes a nucleotide sequence that encodes six consecutive histidine amino acids on the C-terminal end of the target polypeptide for the purpose of protein purification. Target polypeptide is purified by contacting the contents of reaction device with resin modified with Ni<sup>2+</sup> ions. Target polypeptide is eluted from the resin with a solution containing free Ni<sup>2+</sup> ions.

#### Example 12

##### Cellular Production of Target Polypeptides

[0314] Nucleic acids are cloned into DNA plasmids having phage recombination sites and target polypeptides are expressed therefrom in a variety of host cells. Alpha phage genomic DNA contains

short sequences known as attP sites, and *E. coli* genomic DNA contains unique, short sequences known as attB sites. These regions share homology, allowing for integration of phage DNA into *E. coli* via directional, site-specific recombination using the phage protein Int and the *E. coli* protein IHF. Integration produces two new att sites, L and R, which flank the inserted prophage DNA. Phage excision from *E. coli* genomic DNA can also be accomplished using these two proteins with the addition of a second phage protein, Xis. DNA vectors have been produced where the integration/excision process is modified to allow for the directional integration or excision of a target DNA fragment into a backbone vector in a rapid *in vitro* reaction (Gateway™ Technology (Invitrogen, Inc.)).

[0315] A first step is to transfer the nucleic acid insert into a shuttle vector that contains attL sites surrounding the negative selection gene, ccdB (*e.g.* pENTER vector, Invitrogen, Inc.). This transfer process is accomplished by digesting the nucleic acid from a DNA vector used for sequencing, and to ligate it into the multicloning site of the shuttle vector, which will place it between the two attL sites while removing the negative selection gene ccdB. A second method is to amplify the nucleic acid by the polymerase chain reaction (PCR) with primers containing attB sites. The amplified fragment then is integrated into the shuttle vector using Int and IHF. A third method is to utilize a topoisomerase-mediated process, in which the nucleic acid is amplified via PCR using gene-specific primers with the 5' upstream primer containing an additional CACC sequence (*e.g.*, TOPO® expression kit (Invitrogen, Inc.)). In conjunction with Topoisomerase I, the PCR amplified fragment can be cloned into the shuttle vector via the attL sites in the correct orientation.

[0316] Once the nucleic acid is transferred into the shuttle vector, it can be cloned into an expression vector having attR sites. Several vectors containing attR sites for expression of target polypeptide as a native polypeptide, N-fusion polypeptide, and C-fusion polypeptides are commercially available (*e.g.*, pDEST (Invitrogen, Inc.)), and any vector can be converted into an expression vector for receiving a nucleic acid from the shuttle vector by introducing an insert having an attR site flanked by an antibiotic resistant gene for selection using the standard methods described above. Transfer of the nucleic acid from the shuttle vector is accomplished by directional recombination using Int, IHF, and Xis (LR clonase). Then the desired sequence can be transferred to an expression vector by carrying out a one hour incubation at room temperature with Int, IHF, and Xis, a ten minute incubation at 37°C with proteinase K, transforming bacteria and allowing expression for one hour, and then plating on selective media. Generally, 90% cloning efficiency is achieved by this method. Examples of expression vectors are pDEST 14 bacterial expression vector with att7 promoter, pDEST 15 bacterial expression vector with a T7 promoter and a N-terminal GST tag, pDEST 17 bacterial vector with a T7 promoter and a N-terminal polyhistidine affinity tag, and pDEST 12.2 mammalian expression vector with a CMV promoter and neo resistance gene. These expression vectors or others like them are transformed or transfected into

cells for expression of the target polypeptide or polypeptide variants. These expression vectors are often transfected, for example, into murine-transformed adipocyte cell line 3T3-L1, (ATCC), human embryonic kidney cell line 293, and rat cardiomyocyte cell line H9C2.

### Example 13

#### Haplotype analysis of the *KIAA0861* locus

[0317] rs6804951 and rs2001449 are significant at the allele and genotype levels ( $P < 0.05$ ).

Moderate LD is observed for markers rs3732602 and rs2293203 ( $r^2 = 0.646$ ). Chi-squared tests indicate that haplotypes are significantly associated with breast cancer. Cell-specific chi-square values indicate that TTTTG and CTTTC haplotypes are contributors to this relationship. Odds ratios and score tests indicate that individuals carrying the TTTTG are less likely to have breast cancer, while individuals with CTTTC are at elevated risk for the disease. Moreover, the odds ratio estimated for the CGTTC indicates more than a two-fold risk of disease among its carriers, although this result must be interpreted with great caution due to the low observed frequency in the population.

#### A. Summary Statistics of Alleles and Genotypes

##### 1. SNP Locations

SNP.ID	Type	Location
rs6804951	Proximal	184327431
rs7639705	Proximal	184330963
rs3732602	Proximal	184408945
rs2293203	Proximal	184419992
rs2001449	Incident	184429569

##### 2. Allele by GYNGroup

	N	Case (N=544)	Control (N=552)	Test Statistic
rs6804951 : T	1064	5% ( 24)	9% ( 46)	Chi-square=6.71 d.f.=1 P=0.00958
rs7639705 : T	1086	80% (434)	81% (441)	Chi-square=0.03 d.f.=1 P=0.868
rs3732602 : T	1074	99% (532)	99% (532)	Chi-square=0.4 d.f.=1 P=0.529
rs2293203 : T	1088	99% (536)	99% (538)	Chi-square=0.27 d.f.=1 P=0.6
rs2001449 : C	1084	30% (161)	22% (119)	Chi-square=8.49 d.f.=1 P=0.00356

### 3. Genotype by GYNGroup

	N  Case	Control (N=272)	Test (N=276)		Statistic	
	rs6804951 : CC	532	91% (238)		83% (225)	Chi-square=7.13 d.f.=2 P=0.0283
	CT		9% ( 24)		16% ( 44)	
	TT		0% ( 0)		0% ( 1)	
	rs7639705 : GG	543	3% ( 9)		5% ( 14)	Chi-square=2.03 d.f.=2 P=0.362
	GT		33% ( 88)		28% ( 77)	
	TT		64% (173)		67% (182)	
	rs3732602 : TT	537	99% (264)		98% (263)	Chi-square=0.4 d.f.=1 P=0.527
	rs2293203 : TT	544	98% (265)		97% (265)	Chi-square=0.28 d.f.=1 P=0.598
	rs2001449 : GG 542		47% (128)		60% (162)	Chi-square=9.29 d.f.=2 P=0.00961
	GC		46% (125)		37% ( 99)	
	CC		7% ( 18)		4% ( 10)	

### 4. Genotype QC: Test of Hardy-Weinberg Equilibrium

#### a. Cases

	A.freq	D	ChiSq	Pvalue
rs6804951	0.936	-0.002280	0.7870	0.3750
rs7639705	0.807	0.004790	0.5150	0.4730
rs3732602	0.990	-0.000101	0.0565	0.8120
rs2293203	0.987	-0.000164	0.0921	0.7620
rs2001449	0.744	-0.014500	3.1400	0.0763

#### b. Controls

	A.freq	D	ChiSq	Pvalue
rs6804951	0.916	-0.003400	0.5350	0.465
rs7639705	0.808	0.014400	2.3600	0.124
rs3732602	0.989	-0.000120	0.0336	0.855
rs2293203	0.985	-0.000213	0.0601	0.806
rs2001449	0.783	-0.010700	1.0800	0.299

### B. Summary Statistics: Linkage Disequilibrium

#### 1. PHASE Haplotype Frequencies

	H.freq	H.relfreq
CGTTC	13	0.012
CGTTG	191	0.175
CTCAG	10	0.009
CTCTG	1	0.001
CTTAG	4	0.004
CTTTC	265	0.243
CTTTG	538	0.493
TGTTG	7	0.006
TTTTC	2	0.002
TTTTG	61	0.056

## 2. Linkage Disequilibrium Between Markers

### a. $r^2$

	rs6804951	rs7639705	rs3732602	rs2293203	rs2001449
rs6804951	1.000000	0.00382	0.000697	0.00089	0.01860
rs7639705	0.003820	1.00000	0.002440	0.00311	0.04770
rs3732602	0.000697	0.00244	1.000000	0.64600	0.00351
rs2293203	0.000890	0.00311	0.646000	1.00000	0.00448
rs2001449	0.018600	0.04770	0.003510	0.00448	1.00000

### b. $D'$

	rs6804951	rs7639705	rs3732602	rs2293203	rs2001449
rs6804951	1.0000	0.116	0.0685	0.0685	0.306
rs7639705	0.1160	1.000	0.2400	0.2400	0.262
rs3732602	0.0685	0.240	1.0000	0.9080	0.345
rs2293203	0.0685	0.240	0.9080	1.0000	0.345
rs2001449	0.3060	0.262	0.3450	0.3450	1.000

### c. P-value

	rs6804951	rs7639705	rs3732602	rs2293203	rs2001449
rs6804951	1.00e+00	4.12e-02	0.3830	0.3240	6.40e-06
rs7639705	4.12e-02	1.00e+00	0.1030	0.0653	5.41e-13
rs3732602	3.83e-01	1.03e-01	1.0000	0.0000	5.03e-02

rs2293203	3.24e-01	6.53e-02	0.0000	1.0000	2.70e-02
rs2001449	6.40e-06	5.41e-13	0.0503	0.0270	1.00e+00

### 3. Haplotype by GYNGroup

#### a. PHASE Haplotypes (All)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TTTTG	20	1.83	3.55	41	3.75	3.53	0.4782	-0.7377
CTCAG	4	0.37	0.19	6	0.55	0.19	0.6654	-0.4074
TGTTG	3	0.27	0.07	4	0.37	0.07	0.7493	-0.2886
CTTTG	259	23.72	0.30	279	25.55	0.30	0.9060	-0.0987
CGTTG	94	8.61	0.01	97	8.88	0.01	0.9662	-0.0344
CTTAG	2	0.18	0.00	2	0.18	0.00	1.0000	0.0000
TTTTTC	1	0.09	0.00	1	0.09	0.00	1.0000	0.0000
CTTTC	151	13.83	2.73	114	10.44	2.71	1.3766	0.3196
CGTTC	9	0.82	0.98	4	0.37	0.98	2.2604	0.8155
CTCTG	1	0.09	0.51	0	0.00	0.50	Inf	Inf

Pearson Chi-squared Test = 16.6377, DF = 9, P-value = 0.0547

#### b. PHASE Haplotypes (Low Frequency Removed)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TTTTG	20	1.86	3.55	41	3.80	3.52	0.4781	-0.7379
CTCAG	4	0.37	0.19	6	0.56	0.19	0.6654	-0.4074
CTTTG	259	24.03	0.30	279	25.88	0.30	0.9056	-0.0992
CGTTG	94	8.72	0.01	97	9.00	0.01	0.9661	-0.0345
CTTTC	151	14.01	2.73	114	10.58	2.71	1.3774	0.3202
CGTTC	9	0.83	0.98	4	0.37	0.98	2.2605	0.8156

Pearson Chi-squared Test = 15.4946, DF = 5, P-value = 0.008445

#### c. haplo.score Haplotypes

	Hap.Freq	Score	P. X^2	P.Sim
TTTTG	0.0529	-2.1206	0.0340	0.0342

TGTTG	0.0101	-2.0668	0.0388	0.0236
CTCAG	0.0073	-1.2914	0.1966	0.2902
CTTTG	0.5221	-1.2275	0.2196	0.2195
CGTTG	0.1448	-0.1441	0.8854	0.8834
CTTTC	0.2267	2.3422	0.0192	0.0192
CGTTC	0.0307	2.6994	0.0069	0.0050

Global Score = 20.343, DF = 7, Global  $P.X^2$  = 0.0049, Global P.Sim = 0.0022

#### Example 14

##### Haplotype analysis of the *NUMA1* locus

[0318] All markers noted below except 2276396 are associated with breast cancer at the allele level ( $P < 0.05$ ). Marker 675185 does not maintain this relationship at the genotype level. Strong LD is observed across the entire region but is particularly strong between and among 1894003, 675185, 673478, and 615000. Pearson chi-squared statistics suggest that haplotypes are significantly associated with breast cancer. Haplotype TTCTC contributes the most to this relationship. Odds ratios and score statistics indicate that individuals with haplotype TTCTC are 2.6 times more likely to have breast cancer than individuals with other haplotypes.

#### Statistics

[0319] Chi-squared statistics are estimated to assess whether 1) alleles and genotypes are associated with breast cancer status and 2) marker genotype frequencies deviate significantly from Hardy-Weinberg equilibrium (HWE). Haplotype frequencies and relative frequencies are estimated, as well as several statistics ( $r^2$ ,  $D'$ , and p-value) that gauge the extent and stability of linkage disequilibrium between markers in each region. Chi-squared statistics and score tests are estimated to determine whether reconstructed haplotypes are significantly associated with breast cancer status ( $P < 0.05$ ). P-values are estimated for 1) the full set of reconstructed haplotypes and 2) a reduced set that excludes haplotypes with observed frequencies less than 10. Results are presented by chromosome order.

## Results

### Summary Statistics: Alleles and Genotypes

#### SNP Locations

SNP.ID	Type	Location
1894003	Proximal	71972974
675185	Proximal	71998270
673478	Incident	72021802
615000	Proximal	72025650
2276396	Proximal	72046603

#### Allele by GYNGroup

	N	Case(N=510)	Control (N=538)	Test Statistic
1894003:C	1026	91%(450)	96%(510)	Chi-square=6.95 d.f.=1 P=0.00838
675185:G	1010	92%(451)	95%(498)	Chi-square=3.96 d.f.=1 P=0.0466
673478:C	1022	8%(41)	5%(25)	Chi-square=5.68 d.f.=1 P=0.0171
615000:G	1010	92%(434)	96%(513)	Chi-square=7.4 d.f.=1 P=0.00652
2276396:C	1028	97%(478)	98%(523)	Chi-square=0.18 d.f.=1 P=0.674

#### Genotype by GYNGroup

	N	Case (N=255)	Control (N=269)	Test Statistic
1894003:TT	513	1%(3)	0%(0)	Chi-square=7.43 d.f.=2 P=0.0243
TC		15%(36)	9%(24)	
CC		84%(207)	91%(243)	
675185:TT	505	0%(1)	0%(0)	Chi-square=4.37 d.f.=2 P=0.112
TG		14%(35)	9%(24)	
GG		85%(208)	91%(237)	
673478:TT	511	84%(207)	91%(241)	Chi-square=6.39 d.f.=2 P=0.0409
TC		14%(35)	9%(25)	
CC		1%(3)	0%(0)	
615000:TT	505	1%(3)	0%(0)	Chi-square=7.8 d.f.=2 P=0.0202



TG		14%(34)	9%(23)	
GG		84%(200)	91%(245)	
2276396:CC	514	4%(232)	95%(255)	Chi-square=0.18 d.f.=1 P=0.67

**Genotype QC: Test of Hardy-Weinberg Proportions**

**All**

	<b>A.freq</b>	<b>D</b>	<b>ChiSq</b>	<b>Pvalue</b>
1894003	0.935	0.00159	0.350	0.554
675185	0.935	0.00159	0.350	0.554
673478	0.935	0.00159	0.350	0.554
615000	0.937	0.00184	0.495	0.482
2276396	0.974	-0.00069	0.374	0.541

**Control**

	<b>A.freq</b>	<b>D</b>	<b>ChiSq</b>	<b>Pvalue</b>
1894003	0.953	-0.002190	0.644	0.422
675185	0.953	-0.002190	0.644	0.422
673478	0.953	-0.002190	0.644	0.422
615000	0.957	-0.001860	0.541	0.462
2276396	0.976	-0.000593	0.166	0.683

**Summary Statistics: Linkage Disequilibrium**

**Haplotype Frequencies**

	<b>H.freq</b>	<b>H.relfreq</b>
CGTGC	961	0.935
TTCGC	1	0.001

TTCGG	1	0.001
TTCTC	39	0.038
TTCTG	26	0.025

Linkage Disequilibrium Between Markers

$r^2$

	1894003	675185	GP11.079035103	615000	2276396
1894003	1.000	1.000	1.000	0.968	0.387
675185	1.000	1.000	1.000	0.968	0.387
673478	1.000	1.000	1.000	0.968	0.387
615000	0.968	0.968	0.968	1.000	0.369
2276396	0.387	0.387	0.387	0.369	1.000

$D'$

	1894003	675185	GP11.079035103	615000	2276396
1894003	1	1	1	1.00	1.00
675185	1	1	1	1.00	1.00
673478	1	1	1	1.00	1.00
615000	1	1	1	1.00	0.96
2276396	1	1	1	0.96	1.00

**P-value**

<b>X</b>	1894003	675185	GP11.079035103	615000	2276396
1894003	1	0	0	0	0
675185	0	1	0	0	0
GP11.079035103	0	0	1	0	0
615000	0	0	0	1	0
2276396	0	0	0	0	1

### Haplotype by GYNGroup

#### PHASE Haplotypes (All)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
TTCGC	0	0.00	0.48	1	0.10	0.44	0.0000	-Inf
TTCGG	0	0.00	0.48	1	0.10	0.44	0.0000	-Inf
CGTGC	452	43.97	0.21	509	49.51	0.19	0.8001	-0.2230
TTCTG	14	1.36	0.18	12	1.17	0.17	1.1690	0.1561
TTCTC	28	2.72	4.57	11	1.07	4.23	2.5887	0.9512

Pearson Chi-squared Test = 11.4058, DF = 4, P-value = 0.02236

Permutation Test P-value = 0.14

#### PHASE Haplotypes (Low Frequency Excluded)

	Case	Case(%)	Case.X^2	Control	Control(%)	Control.X^2	OR	ln.OR
CGTGC	452	44.05	0.25	509	49.61	0.23	0.7998	-0.2234
TTCTG	14	1.36	0.18	12	1.17	0.16	1.1690	0.1561
TTCTC	28	2.73	4.53	11	1.07	4.21	2.5888	0.9512

Pearson Chi-squared Test = 9.5506, DF = 2, P-value = 0.008435

#### haplo.score Haplotypes

	Hap.Freq	Score	P.X^2	P.Sim
CGTGC	0.9410	-2.0316	0.0422	0.0531
TTCTG	0.0248	0.3232	0.7465	0.8344
TTCTC	0.0321	2.6973	0.0070	0.0093

Global Score = 9.1386, DF = 3, Global P.X^2 = 0.0275, Global P.Sim = 0.0212

[0320] Modifications may be made to the foregoing without departing from the basic aspects of the invention. Although the invention has been described in substantial detail with reference to one or more specific embodiments, those of skill in the art will recognize that changes may be made to the embodiments specifically disclosed in this application, yet these modifications and improvements are within the scope and spirit of the invention, as set forth in the claims which follow. All publications or patent documents cited in this specification are incorporated herein by reference as if each such publication or document was specifically and individually indicated to be incorporated herein by reference.

[0321] Citation of the above publications or documents is not intended as an admission that any of the foregoing is pertinent prior art, nor does it constitute any admission as to the contents or date of these

publications or documents. U.S. patents, documents and other publications referenced herein are hereby incorporated by reference.